LOW–RATE ANALYSIS–BY–SYNTHESIS WIDEBAND SPEECH CODING

by

Guylain Roy
B. Eng.

Department of Electrical Engineering
McGill University
Montreal, Canada

August 1990

A thesis submitted to the Faculty of Graduate
Studies and Research in partial fulfillment of
the requirements for the degree of
Master of Engineering

Abstract

This thesis studies low–rate wideband analysis–by–synthesis speech coders. The wideband speech signals have a bandwidth of up to 8 kHz and are sampled at 16 kHz, while the target operating bit rate is 16 kbits/sec. Applications for such a coder range from high-quality voice–mail services to teleconferencing. In order to achieve a low operating rate, the coding places more emphasis on the lower frequencies (0 to 4 kHz), while the higher frequencies (4 to 8 kHz) are coded less precisely but with little perceived degradation.

The study consists of three stages. First, aspects of wideband spectral envelope modeling using Line Spectral Frequencies (LSF's) are studied. Then, the underlying coder structure is derived from a basic Residual Excited Linear Predictive coder (RELP). This structure is enhanced by the addition of a pitch prediction stage, and by the development of full–band and split–band pitch parameter optimization procedures. These procedures are then applied to an Code Excited Linear Prediction (CELP) model. Finally, the performance of full–band and split–band CELP structures are compared.

Sommaire

Cette thèse présente une étude de codeurs de parole de large bande et opérant à faible débit. Ces codeurs, du type d'analyse–par–synthèse, doivent transmettre des signaux possédant une gamme de fréquences limitée à 8 kHz et ayant été échantillonés à 16 kHz. Le débit visé est de 16 kbits/sec. Les applications pratiques incluent, entres autres, les services de messagerie de parole de haute qualité ansi que les services de téléconférences. Afin d'obtenir un faible débit, l'emphase est mise sur les basses fréquences (0 à 4 kHz). Les hautes fréquences (4 à 8 kHz) sont, pour leur part, codées de façon moins précise, tout en respectant un faible niveau de distorsion.

L'étude est répartie en trois étapes. En premier lieu, le codage de l'envelope spectrale d'un signal de large bande est étudié à l'aide des Lignes de Fréquences Spectrales. Ensuite, une structure fondée sur le codeur de Prédiction Linéaire à Excitation Résiduelle est mise en place. Cette structure est améliorée par l'addition d'un module de prédiction de la fréquence fondamentale. Des procédures d'optimisation, applicables sur toute la bande passante du signal ou sur des bandes séparées sont ensuite développées. Ces procédures sont finalement adaptées à un modèle de Prédiction Linéaire à Excitation Vectorielle, et leurs performances sont analysées.

Acknowledgments

I would like to thank my supervisor, Dr. Peter Kabal, for his guidance during this study, and for his financial assistance. The facilities and the professional environment provided by INRS-Télécommunications were most appreciated.

Also, I cannot overlook the healthy lighter moments sparked by the combined wits of Duncan, Daniel and Ravi. Finally, I am most grateful to my parents, Palme and Yvan, for their constant love and support, and to Katrina, for her patience.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

The main goal of speech coding is to efficiently transform an analog voice waveform into a digital bit stream. This bit stream can either be transmitted over a digital channel (e.g. telephone), or simply stored for later playback (e.g. voice–mail). In both cases, the coding scheme is subjected to two related factors: 1) the desired reproduced speech quality (usually a function of the end user application), and 2) the operating bit rate (which depends on the transmission channel or storage medium capacity). Thus, the coding efforts can be channeled towards improving the reproduced speech quality given an operating bit rate, or towards reducing the bit rate while preserving an acceptable reproduced speech quality.

Over the years, proposed coding techniques have covered a wide range of operating bit rates for an equivalently wide range of reproduced speech quality. This reproduced speech quality is a direct function of the speech signal bandwidth. With time, two main categories have emerged: *narrowband* speech coding (also known as *digital telephony*), and *wideband* speech coding (a subset of *digital audio*).

1.1 Narrowband and Wideband Speech Coding

In narrowband digital speech systems, the speech bandwidth is limited to 3.4 kHz and the sampling rate is set at 8 kHz. Applications, other than telephony, include mobile radio, voice mail and secure voice communications. Coding standards exist for rates of 64 kbits/sec all the way down to 2.4 kbits/sec. The coding algorithms vary from the high rate/low complexity waveform coders to the medium to low rate/high complexity vocoders or hybrid coders [1]. In their simplest form, waveform coders simply process each incoming speech sample independently. On the otherhand, vocoders try to fit the incoming speech to a set of pre-determined parameters based on well known speech production models. Finally, hybrid coders, as their name implies, are a combination of waveform coders and vocoders.

These coding algorithms can also be grouped into three quality categories: synthetic, communications and toll (or network) quality [1]. Vocoders operating at rates less than, or equal to 2.4 kbits/sec generally produce synthetic speech quality. In this case, the speech is intelligible but is generally machine–like and stripped of much of its naturalness. Communications quality speech is generally produced by hybrid coders operating at rates of 4.8 to 9.6 kbits/sec. The coded speech usually sounds natural, although it still exhibits some noticeable distortion. Finally, toll quality speech is produced by waveform coders operating at 16 to 64 kbits/sec. Telephone speech falls into this quality category.

In recent years, most narrowband speech coding research efforts have been focused on communications or toll quality coders operating at, or below 16 kbits/sec. In

particular, successful implementations of CELP (Code Excited Linear Predictive) coders [2] have been shown to yield communications quality speech for rates as low as 4.8 kbits/sec.

In wideband speech coding, the speech bandwidth varies from 7 kHz to 20 kHz. The three main quality categories are commentary (also known as AM radio), FM radio and compact disc. These usually require bit rates ranging from 64 kbits/sec all the way up to 700 kbits/sec, for sampling rates varying between 16 kHz and 45 kHz [3]. From a digital audio perspective, these categories also include music signals which may have spectral components up to 20 kHz. However, 7 to 10 kHz is sufficient to represent nearly all of the perceivable speech information.

The coding of 7 kHz speech is of particular interest since it offers a substantial increase in perceived quality, yet at rates similar to those found in high quality narrowband telephony systems. In particular, the CCITT (Consultative Committee for Telephone and Telegraph) standard for 7 kHz audio (G.722) operates at 64 kbits/sec and is primarily intended for the ISDN (Integrated Service Digital Network) environment. As is, the G.722 standard offers twice the speech bandwidth with the same bit rate required for simple coding of narrowband telephone speech.

1.2 Motivation

The narrowing of the gap between digital telephony and digital audio operating bit rates leads one believe in the possible implementation of wideband speech coders operating at lower bit rates. From now on, the term *wideband* in this research will be

limited to speech with a 7.5 kHz bandwidth and sampled at 16 kHz. Since roughly 80% of the perceptually important speech spectral information is contained within the baseband (0.2 to 3.2 kHz) [1], it is reasonable that the incremental cost of coding the extra bandwidth of a wideband signal should be relatively small. The added bandwidth yields a fuller, richer sound and the high frequency spectral content helps differentiate among fricatives (e.g. *"s"* versus *"f"*).

This thesis is a study of possible implementations for a low–rate wideband coder. The goal is to produce a system capable of coding wideband speech at rates equal to, or less than 16 kbits/sec. This ceiling rate is based on the successes of coding for digital telephony at 9.6 kbits/sec. The structures being investigated are known as *analysis–by–synthesis* coders. These *smart* coders reproduce and optimize the coded speech before transmission. This allows for an iterative and/or joint selection of the coding parameters that yield the best possible coded speech under the constraints of parameter quantization. CELP coders fall into this category, which lies somewhere between standard waveform coders and the general Linear Predictive Coding (LPC) category. These structures are reviewed in Chapter 2.

Two fundamental approaches are taken for this study: a *full–band* implementation and a *split–band* implementation. In the full–band case, the input speech signal is analyzed and coded with all of its frequency contents considered together. In the split–band approach, the low (0 to 4 kHz) and high bands (4 to 8 kHz) of the input speech signal are dealt with separately. This provides flexible control over the coding resolution given to the low and high frequency components of the speech. To help study both approaches, a basic hybrid coder structure, known as RELP

(Residual Excited Linear Prediction), is used as a development platform. The basic RELP coder is also presented in Chapter 2. This coder, used for the first time in a wideband context, helps determine how the high band of the speech should be coded. In particular, it will help demonstrate how the high frequency components of the speech can be sub-optimally reproduced with very little, if not negligible perceived distortion, provided that the low frequency components are very well coded.

1.3 Scope and Organization of the Thesis

This thesis contains six chapters. The second chapter contains background material such as LPC techniques, RELP and CELP coder structures.

The third chapter is a study of wideband spectral envelope coding. In particular, the behavior of Line Spectral Frequencies (LSF's) in a wideband context is analyzed, and various LSF coding methods are comparatively tested.

The fourth chapter deals with RELP in more detail. First, the basic RELP model is enhanced by the addition of a pitch prediction loop. This model is then transformed into an analysis–by–synthesis structure by the development of pitch optimization procedures. These procedures are flexible and either operate in full or split–band mode. Both modes are studied with no quantization to determine the effects residual band-limiting and ways of coding the high frequencies of the signal. Finally, aspects of efficient residual coding are covered and migration to CELP justified.

The fifth chapter presents wideband CELP models. The basic analysis–by–synthesis RELP models of Chapter 4 are transformed into full and split–band CELP

structures. Various aspects parameter selection and coding are investigated for both structures. In particular, the effects of band-limiting the codebooks are studied for the split–band structure. Also, an efficient sub-optimal high–band codeword selection scheme is presented for the split–band structure. Finally, it is found that the split–band approach is superior as both structures are compared while subjected to a maximum operating rate of 16 kbits/sec.

The last chapter concludes with a summary of the results. Suggestions and directions for future research are proposed.

Chapter 2 Background Material

This chapter covers background material used in this research. It first reviews basic LPC (Linear Prediction Coding) systems, and then describes the more advanced RELP (Residual Excited Linear Prediction) and CELP systems.

2.1 Linear Prediction Coding (LPC)

A Linear Predictive Coding system attempts to extract a set of parameters that best describes the signal it analyzes. In particular, it assumes that the signal has been produced by a source exciting one or more linear filters. Although this is an imprecise model for real life signals (speech, radar, seismic signals), it nevertheless constitutes a good estimate of how these signals are produced. This allows for a compact parametric representation of any signal that match the linear production model. In turn, this translates into more efficient transmission systems where the model parameters, rather than signal itself, are coded and sent. For human speech, there are two common types of linear prediction: *short–term* (also known as *formant*) and *long–term* (also known as *pitch*) prediction.

2.1.1 Short–term prediction

In short–term prediction, the linear production model assumes that the excitation source is spectrally flat. This leaves all the spectral shaping to the linear filter. This filter corresponds to the vocal tract shape and size and is usually referred to as the *formant synthesis* filter. Since the vocal tract changes slowly with time, speech is considered to be stationary within finite small time intervals (*frames*). The filter coefficients can thus be updated only every 10 to 20 ms and still closely match the signal spectral behaviour.

A general formant synthesis filter has P poles and Q zeros. However, in most speech LPC analysis techniques, the formant synthesis filter $H(z)$ is assumed to be an *all–pole* filter (i.e. Q=0). This greatly simplifies the derivation of its parameters by reducing the LPC analysis to solving a set of linear equations. The all–pole model exhibits resonances which are well suited for representing the spectral peaks found in speech. However, the spectral valleys, which correspond to zeros, cannot be accurately modeled. Fortunately, the human ear is much more sensitive to spectral peaks than to valleys and the all–pole model remains a well–founded approach. The number of poles P is a function of the number of formants to be modeled. Generally, each formant requires two poles, while two extra poles are added to compensate for the glottal effects and radiation at the lips [1]. In digital telephony, this translates to 10 poles, whereas wideband speech is best represented by 16 or more poles. Also, to ensure the stability of the synthesis filter, all the poles must lie inside the unit circle.

Let $s(n)$ be a discrete speech signal obtained from sampling $s(t)$ at a rate of F_s.

Based on the suggested production model, let the formant prediction filter $F(z)$ be defined as

$$F(z) = \sum_{k=1}^{P} a_k z^{-k} \tag{2.1}$$

where the a_k are the LPC coefficients.

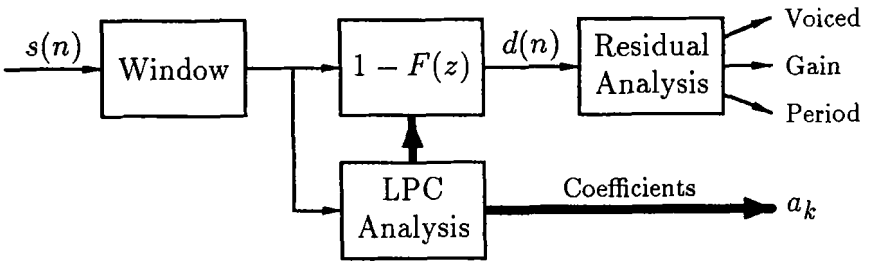
When the input speech $s(n)$ is passed through the *inverse formant* (or *formant error prediction*) filter $A(z) = 1 - F(z)$, the resulting error signal $d(n)$ is:

$$d(n) = s(n) - \sum_{k=1}^{P} a_k s(n-k). \tag{2.2}$$

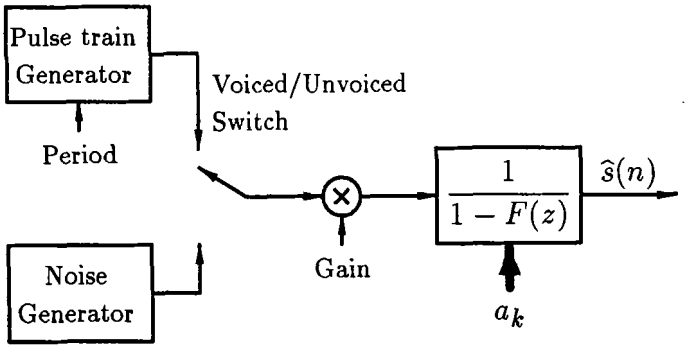
The error signal or *formant residual* $d(n)$ can be viewed as a speech signal from which linear near–sample redundancies have been removed. It is very noise–like during unvoiced segments of speech, while voiced sections show a clear embedded pulse train structure (also referred to as *pitch* or *fine line* structure). The distance between each pulse is the pitch period and corresponds to the regular puffs of air passing through the vibrating vocal folds of the glottis.

The LPC coefficients a_k are derived so as to minimize, in the mean-square sense, the error $d(n)$ over the analysis frame. The *autocorrelation* and *covariance* methods are well known least-squares techniques used to find the a_k [4]. The autocorrelation method always yields a set of stable LPC coefficients (i.e. all the poles of the formant synthesis filter lie inside the unit circle). This is not always the case for the covariance method. Furthermore, it should be noted that the LPC coefficient stability may no longer be preserved after quantization. This can happen when the coefficient coding scheme used cannot accommodate the wide dynamic ranges of the a_k. To counter this effect, more efficient alternate LPC representations, such as reflection coefficients,

log-area ratios and Line Spectral Frequencies, have been developed [1]. These usually have better quantization properties than the direct form coefficients a_k, and can also provide for stability checks. The Line Spectral Frequencies are studied in more detail in Chapter 3.



(a) analysis phase



(b) synthesis phase

Fig. 2.1 Basic LPC vocoder.

A simple LPC vocoder structure is shown in Figure 2.1. Rather than being transmitted, the error signal $d(n)$ is modeled by three parameters: a gain factor G, a voiced/unvoiced decision, and a pitch period estimate for voiced sections. These are sent along with the LPC coefficients and are used at the receiver to regenerate $\hat{s}(n)$

by exciting the all–pole formant synthesis filter $H(z)$:

$$H(z) = \frac{1}{1 - F(z)} \tag{2.3}$$

The resulting speech is usually intelligible but is of synthetic quality. This is a direct consequence of the simple assumptions governing the linear production model. Although noise–like, the error signal $d(n)$ still contains a wealth of information affecting the naturalness and quality of the speech. Limiting its representation to 3 parameters directly reduces the overall quality of the reconstructed speech. Nevertheless, LPC vocoders have found their way into applications requiring very low transmission rates (i.e. less than 2.4 kbits/sec), in particular for secure communications.

2.1.2 Long–term prediction

Let $P(z)$ be a pitch prediction filter defined as follows:

$$P(z) = \beta_1 z^{-(M-1)} + \beta_2 z^{-M} + \beta_3 z^{-(M+1)}. \tag{2.4}$$

In this form, $P(z)$ can be viewed as a 3 tap FIR filter centered around a delay of M samples. The delay M, or *pitch lag*, is the estimate of the pitch period and can be obtained, along with the coefficients β_i, through a signal correlation analysis [5]. The optimal lag and pitch coefficients are obtained by performing the analysis over a pre-defined lag range. In general, the pitch lag is allowed to vary between 2.5 ms and 20 ms. For speech sampled at 16 kHz, this translates to pitch delays varying between 41 and 320 samples. Also, single tap pitch predictors are suitable for most applications. However, the sampling period $T_s = 1/F_s$ is not always an integer factor

of the real pitch period. A multiple tap pitch prediction filter acts somewhat as an interpolator and thus yields more precise pitch delay estimates.

When the formant residual signal $d(n)$ is passed through the *pitch inverse* or *pitch error prediction* filter $B(z) = 1 - P(z)$, the resulting error signal $r(n)$ is defined as:

$$r(n) = d(n) - \beta_1 d(n - M + 1) - \beta_2 d(n - M) - \beta_3 d(n - M - 1). \qquad (2.5)$$

In contrast with formant prediction, long–term prediction attempts to remove far–sample redundancies. The pitch prediction filter estimates the pitch period of the glottal excitation. For unvoiced speech segments, no clear pitch period exists, and the pitch filter is effectively disabled. During voiced speech, the pitch prediction filter removes the pulse train structure from the formant residual signal. A good example of the differences between $d(n)$ and $r(n)$ is shown in Figure 2.2. The pitch residual signal in the bottom trace is obtained with a 3 tap pitch prediction filter. Frames 184 and 185 correspond to a voiced/unvoiced transition. During voiced frames, the pitch structure is visible in the formant residual but is absent from the pitch residual. Both residuals are similar during unvoiced frames.

The use of long–term prediction helps reduce the variance of the prediction residual. In some hybrid coders (e.g. Adaptive Predictive Coding), both formant and pitch prediction are used, and the residual is quantized and sent along with all the coefficients. Quantizing $r(n)$ is more efficient than quantizing $s(n)$ or $d(n)$. This results in better quality reconstructed speech. At the receiver, the quantized pitch residual $\hat{r}(n)$ excites the pitch synthesis filter $G(z)$ defined as:

$$G(z) = \frac{1}{1 - P(z)}. \qquad (2.6)$$

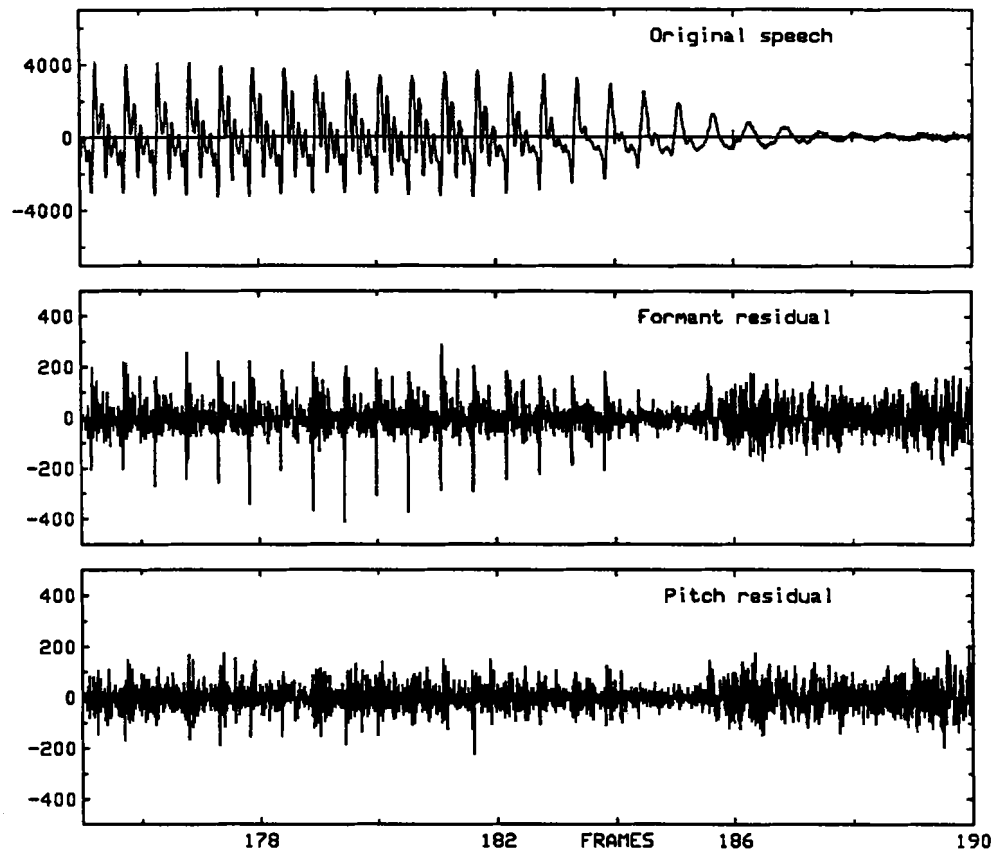


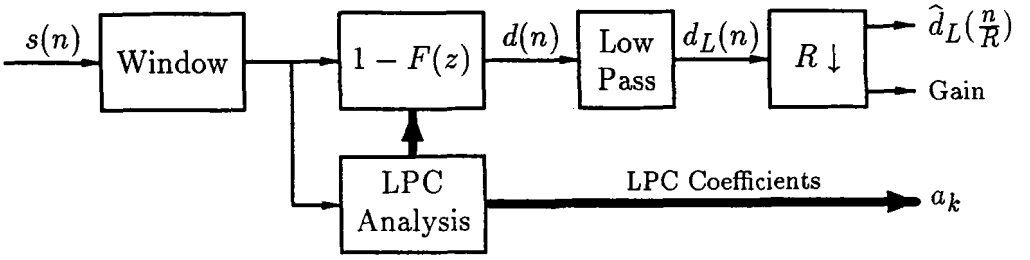
Fig. 2.2 Comparison of formant and pitch residuals.

The output of $G(z)$ corresponds to a quantized formant residual, which in turn excites the formant synthesis filter $H(z)$.

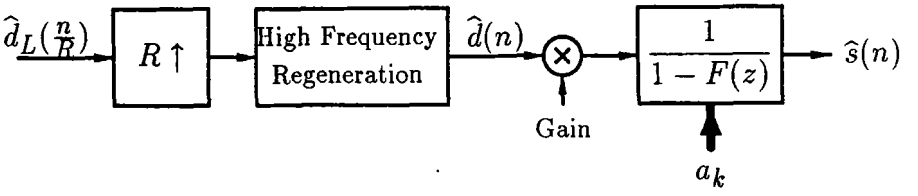
Finally, because it is an IIR filter, the pitch synthesis filter is subject to instability. During unvoiced speech segments, or at the onset of voiced segments, the pitch coefficients can lead to an unstable synthesis filter. This has been studied by Ramachandran [5]. A simple stabilization method has been developed for systems having up to 5 pitch coefficients.

2.2 Residual Excited Linear Prediction Coding (RELP)

The original RELP configuration, suggested by Un and Magill [6] can be considered one of the first hybrid coders. As opposed to the basic LPC vocoder system, the RELP coder sends part of the original formant residual signal to the receiver. The excitation signal is thus more realistic, and the reproduced speech of better quality. The basic RELP coder structure is shown, simplified, in Figure 2.3.



(a) analysis phase



(b) synthesis phase

Fig. 2.3 Basic RELP coder.

Let $s(n)$ be a discrete time speech signal with sampling frequency F_s. The formant residual $d(n)$, obtained from a standard LPC analysis, is low-pass filtered. The upper band is simply discarded. The baseband residual signal is then decimated by an

integer factor R, quantized and sent to the receiver, along with the LPC coefficients a_k. For narrowband speech, the decimation ratio is around 4, yielding only 1000 Hz of original baseband information.

At the receiver, the baseband residual is upsampled by R, and a High Frequency Regeneration (HFR) scheme is used to artificially recreate the discarded upper band residual. The regenerated residual $\widehat{d}(n)$ then excites the formant synthesis filter.

The quality of RELP coded speech strongly depends on the HFR scheme used. The formant residual signal $d(n)$ is not always perfectly spectrally flat and can contain a harmonic structure. In those cases, the HFR scheme must then recreate a uniform pitch structure across the whole band. Various HFR methods have been proposed: spectral folding, spectral translation, non-linear functions and hybrid combinations [7].

In the case of spectral HFR methods, problems occur at the boundaries between replicated basebands. In particular, discontinuities in the pitch structure introduce tonal noises, giving the reproduced speech a metallic sound. On the other hand, non-linear functions, such as absolute value, squaring or clipping can regenerate a uniform pitch structure. The resulting residual spectrum however, is never quite flat, and other methods, such as spectral tilt, must be used to compensate these side effects. Finally, hybrid methods make use of both spectral and non-linear approaches. No one method is clearly better than the others and all have been used with some degree of success for medium quality coders, operating in the 4.8 to 9.6 kbits/sec range.

2.3 Code Excited Linear Prediction Coding (CELP)

CELP coding falls in the *analysis–by–synthesis* category of linear predictive systems. These coders offer a full parametric representation of speech signals, and can produce communications quality output at rates as low as 4.8 kbits/sec [2,8]. The term *analysis–by–synthesis* means that the speech coding analysis is done at the transmitter by synthesizing speech signals using pre-determined synthesis parameters (i.e. lag values, quantized pitch coefficients and quantized residual waveforms). The synthesis parameters that yield the best match between the original and coded speech signals are sent to the receiver.

In CELP coders, since both formant and pitch prediction are used, the residual excitation signal is noise–like. CELP coders often exploit this by viewing it as Gaussian noise. The residual waveform is coded using B bits pointing to an entry in a codebook of 2^B waveforms. A simple CELP coder structure is shown in Figure 2.4.

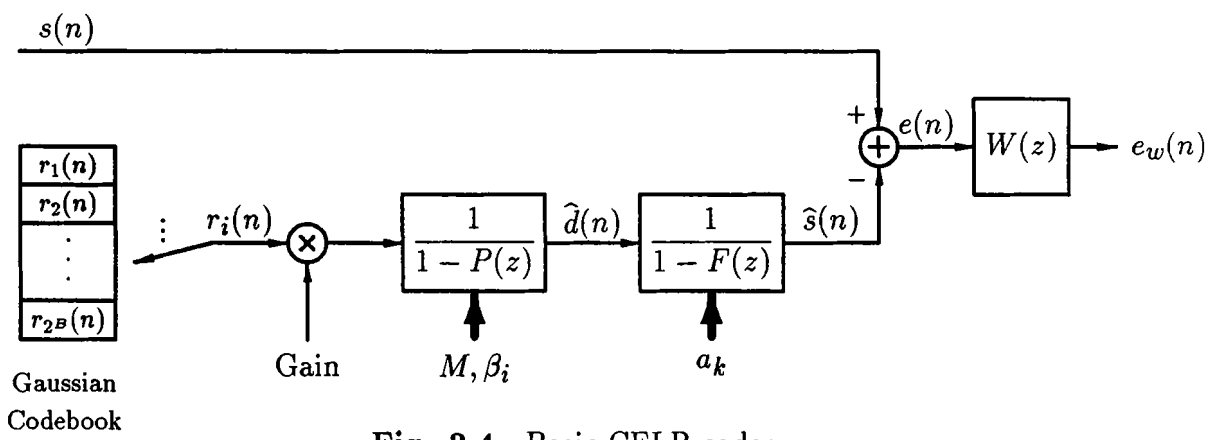


Fig. 2.4 Basic CELP coder.

An LPC analysis is first used to obtain the LPC coefficients. At the synthesis stage, the formant frame (e.g. 20 ms) is further divided into pitch sub-frames (e.g. 5 ms). For each sub-frame, the parameter selection is performed by scanning the codebook, one waveform at a time. For each waveform, the gain G, the pitch lag M and the pitch coefficients β_i are computed such that the weighted error signal $e_w(n)$, defined below, is minimized in the mean-square sense. The index of the waveform yielding the smallest error energy is sent to the receiver, along with the other synthesis parameters. The signal $\hat{s}(n)$ is obtained by exciting the cascaded pitch and formant synthesis filter with the scaled selected waveform (lower branch of Fig 2.4).

The weighted error signal $e_w(n)$ is obtained by passing the error

$$e(n) = s(n) - \hat{s}(n)$$

through the error weighting filter $W(z)$, defined as:

$$\begin{aligned} W(z) &= \frac{H(\gamma z)}{H(z)} \\ &= \frac{1 - F(z)}{1 - F(\gamma z)}. \end{aligned} \tag{2.7}$$

where the bandwidth expansion factor, $\gamma = 1/0.75$, effectively concentrates the coding noise in the formant regions where it is not as perceptible [9].

The resulting speech quality is a function of the codebook size and parameter selection. Codebooks containing as little as 32 waveforms can yield communications quality coded speech. From a practical standpoint, a fully optimal parameter selection is not possible. In particular, the LPC coefficients cannot be easily optimized, due to the low-delay feedback of the formant synthesis filter. They must remain as derived at the analysis stage. However, the pitch parameters (gain, lag and coefficients) can

be re-optimized. When the sub-frame size is smaller than the pitch lag, the pitch lag in the pitch synthesis feedback loop reduces the optimization to solving a set of linear equations. Yet, the computational burden remains heavy. For example, an exhaustive search conducted for a system with a 32 codewords, 128 possible lag values, one pitch coefficient with 16 possible quantized values and a sub-frame size of 5 ms translates to over 13 million optimization iterations per second. This is not practical and thus, more efficient parameter selection methods must be introduced. These will be further discussed in Chapter 5.

Chapter 3

Wideband Spectral Envelope Coding

In this research, Line Spectral Frequencies (LSF's) are used for transmitting the wideband spectral envelope information. LSF's are recognized as one of the most efficient representation for short–time speech coding in low bit rate applications. LSF's are a mathematical transformation of the standard LPC coefficients, and can be converted to other well known linear predictive coefficient representations (e.g. reflection, cepstral [1]). However, their natural properties simplify the quantization procedures, easily ensure synthesis filter stability, allow frame to frame interpolation and provide flexible spectral distortion control.

Although much work has been done with LSF coding, most studies deal with narrowband applications involving low order systems, typically with 10 poles or so. For those systems, operating at a frame update rate of 50 Hz, the LSF parameters transmission requires about 1800 bits/sec, with some systems achieving rates as low as 1000 bits/sec [10]. For this wideband research, higher order systems are required, typically with 16 to 20 poles. Although a low transmission rate is desirable, quality is the prime factor and thus, rates as high as 3000 bits/sec will have to be considered.

The next section reviews the mathematical derivation and computational considerations of obtaining LSF. In Section 3.2, scalar LSF coding schemes are investigated and results of comparative tests are given in Section 3.3.

3.1 LSF Representation

3.1.1 Mathematical description

Consider the standard LPC inverse filter:

$$A_P(z) = 1 + \sum_{k=1}^{P} a_k z^{-k} \tag{3.1}$$

This filter can also be expressed in lattice form, thereby corresponding to an acoustical tube model of the vocal tract:

$$A_n(z) = A_{n-1}(z) - \kappa_n z^{-P} A_{n-1}(z^{-1}) \qquad \text{for } n = 1 \ldots P \tag{3.2}$$

where κ_n is the reflection coefficient for the n^{th} tube.

The above recurrence can be extended by one stage with a reflection coefficient κ_{P+1} set to $+1$ (complete closure of the glottis), or to -1 (complete opening of the glottis). The extended model is thus completely characterized by the following two polynomials:

$$\begin{aligned} P(z) &= A_{P+1}(z) - z^{-(P+1)} A_{P+1}(z^{-1}) \\ Q(z) &= A_{P+1}(z) + z^{-(P+1)} A_{P+1}(z^{-1}) \end{aligned} \tag{3.3}$$

For a stable system, Soong and Juang [11] proved the following properties for $P(z)$ and $Q(z)$:

- All roots of $P(z)$ and $Q(z)$ lie on the unit circle,
- The roots of $P(z)$ and $Q(z)$ alternate around the unit circle.

In general, the second property is known as the *ordering property* of the LSF's for which the $P+2$ roots of $P(z)$ and $Q(z)$ satisfy the following relation:

$$0 = \omega_0 < \omega_1 < \omega_2 < \ldots < \omega_P < \omega_{P+1} = \pi \tag{3.4}$$

Note that ω_0 and ω_{P+1}, the extra roots induced by the $(P+1)^{\text{st}}$ stage, are implicit LSF's and need not be transmitted. Also note that the ω_i's belong to $P(z)$ for i odd, and to $Q(z)$ for i even. Finally, note that for stable systems, the LPC to LSF transformation is one-to-one. This is indeed true, since, by its very nature, the LSF representation is an extension of the PARCOR (i.e. reflection coefficients) model which itself is one-to-one. Therefore, any ordered set of LSF's will correspond to a stable synthesis filter. This feature is especially useful when quantizing LSF's.

3.1.2 Computational considerations

Finding the ω_i's can be computationally costly, and several methods have been proposed. Kang and Fransen have suggested an iterative approach [12] based on the phase function of the all–pass ratio filter $R(z)$ defined as:

$$R(z) = \frac{z^{-(P+1)}A(z^{-1})}{A(z)} \tag{3.5}$$

The phase function of $R(z)$ is monotonically decreasing and satisfies the relation:

$$\Phi(\omega_i) = i\pi \qquad \text{for } i = 1 \ldots P \tag{3.6}$$

Finding the LSF's is then a matter of searching along the $\omega = [0, \pi]$ range, and isolating values of ω_i such that $\Phi(\omega_i)$ is within a prescribed error ξ of $i\pi$.

Soong and Juang, on the other hand, first removed the implicit LSF's ω_0 and ω_{P+1} by division, and evaluated the resulting polynomials $P'(z)$ and $Q'(z)$ on the unit circle. After applying a Discrete Cosine Transformation on the coefficients of $P'(\omega)$ and $Q'(\omega)$, finding the roots is then a matter of searching along the $\omega = [0, \pi]$ range and iteratively isolating each ω_i by monitoring sign changes in both polynomials. Finally, a somewhat similar method, proposed by Kabal and Ramachandran [13], maps the upper half of the unit circle to the $[-1, 1]$ range on the real axis through the $x = \cos(\omega)$ transformation. Based on this, $P'(\omega)$ and $Q'(\omega)$ are transformed to $P'(x)$ and $Q'(x)$ using Chebyshev polynomials. Finding the P roots x_i is an iterative interpolated search within subintervals. The LSF's are then obtained by the inverse mapping $\omega_i = \arccos(x_i)$.

In this research, the last method is used to compute the LSF's. This approach is numerically stable and also reduces the computational load by reducing the number of direct trigonometric function evaluations.

3.2 LSF Quantization

In speech coding applications, efficient parameter quantization is essential. Many quantization procedures have already been proposed for encoding LSF in narrowband systems [10,14,15]. Whether scalar quantization or vector quantization is used, the ultimate goal is to preserve the LSF ordering property and minimize the quantization distortion.

For this wideband research, the study is limited to scalar quantization. The three methods investigated are simple and based on some known narrowband LSF

quantization procedures. They are:

- non–uniform quantization (NUQ),
- differential non–uniform quantization (DNUQ),
- time differential non–uniform quantization (TDNUQ).

The study is conducted for system orders of 16, 18 and 20 poles. In all cases, a training set of 4200 non–silent speech frames from the wideband speech database described in Appendix A is used.

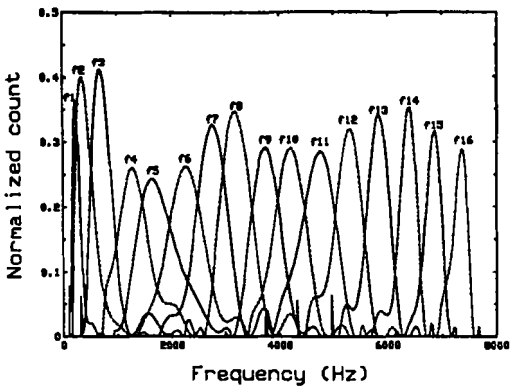
3.2.1 Non–uniform quantization (NUQ)

In this scheme, each LSF is encoded independently. A non–uniform quantizer design is preferred to a uniform design since it yields a lower average quantization error. Given an LSF parameter ω with a range $[\omega_{\min}, \omega_{\max}]$, the M–level quantizer design problem is to minimize the average square error distortion D defined as:

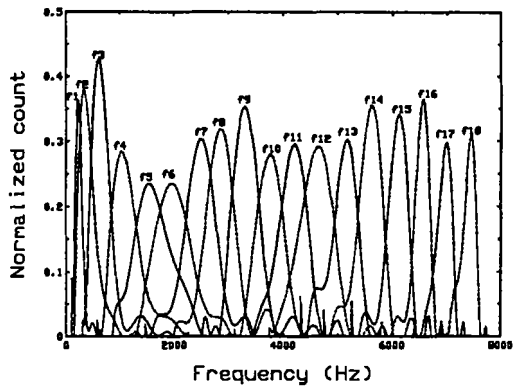
$$\begin{aligned} D &= \int_{\omega_{\min}}^{\omega_{\max}} (\omega - q(\omega))^2 p(\omega) d\omega \\ &= \frac{1}{M} \sum_{j=1}^{M} \int_{T_{j-1}}^{T_j} (\omega - \hat{\omega}_j)^2 p(\omega) d\omega \end{aligned} \tag{3.7}$$

where $q(\omega)$ and $p(\omega)$ are respectively the quantizer and density functions of ω, and T_j and $\hat{\omega}_j$ are the quantizer's decision and output levels.

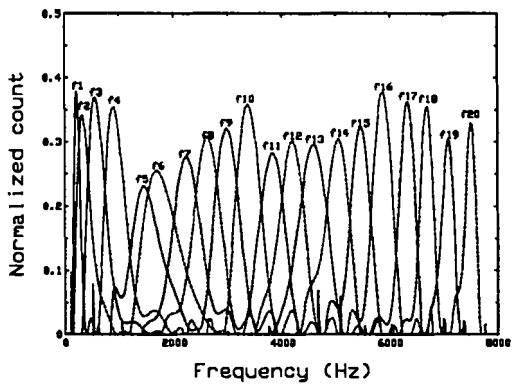
The output levels are chosen to ensure finer resolution for higher probability regions. The quantizers are designed for each LSF parameter using probability densities derived from the training set mentioned earlier. Figure 3.1 shows the overlaid densities for systems with 16, 18 and 20 poles. Note that in all three cases, there is a strong overlap in the 1000 Hz to 3000 Hz range, and that the individual LSF dynamic ranges vary greatly.



(a) 16 poles



(b) 18 poles



(c) 20 poles

Fig. 3.1 LSF densities for 16, 18 and 20 poles.

Because of this overlap, one can expect that after quantization, the ordering property may no longer be respected. This problem is usually limited to adjacent LSF's. A simple remedy is to exchange the values of the adjacent LSF's when this situation is detected at the receiver. However, even though they respect the ordering property, very close quantized LSF's can lead to excessively narrow formant peaks in the coded signal spectral envelope. This problem can be prevented at the receiver by inducing a minimum LSF spacing based on statistical analysis of adjacent LSF spectral distances.

The non–uniform quantization problems mentioned above arise since the quantizer design algorithm only yields a locally optimum structure and has no built-in provision for maintaining the ordering property. Sugamura and Farvardin [14] have proposed a dynamic programming algorithm which yields globally optimum NUQ structure coupled with an algorithm which preserves the ordering property of the LSF. However, this scheme is not used in this research as other simple methods (e.g. DNUQ) are known to perform better.

3.2.2 Differential non–uniform quantization (DNUQ)

Wide and overlapping LSF dynamic ranges complicate the LSF encoding procedures. Soong and Juang [15] proposed a differential LSF encoding method whereby the adjacent spectral distance $d_{i,i+1}$ between neighboring LSF's is encoded rather than the LSF directly. For wideband spectral coding, the statistical analysis shows that the adjacent spectral distance densities possess smaller dynamic ranges (Figure 3.2). These densities are used to design non–uniform quantizers for orders of 16, 18 and 20 poles.

The DNUQ algorithm first encodes ω_1 to $\hat{\omega}_1$ using the NUQ approach described in the previous section. This serves as a reference for encoding subsequent LSF's. Then, for i ranging from 1 to $P-1$, the spectral distance $d_{i,i+1}$ between $\hat{\omega}_i$ and ω_{i+1} is quantized to $\hat{d}_{i,i+1}$. At the receiver, ω_{i+1} is then generated by adding $\hat{\omega}_i$ and $\hat{d}_{i,i+1}$.

The clear advantage of DNUQ over NUQ is that it always preserves the LSF ordering property. However, special care must be taken to ensure the last few LSF's do

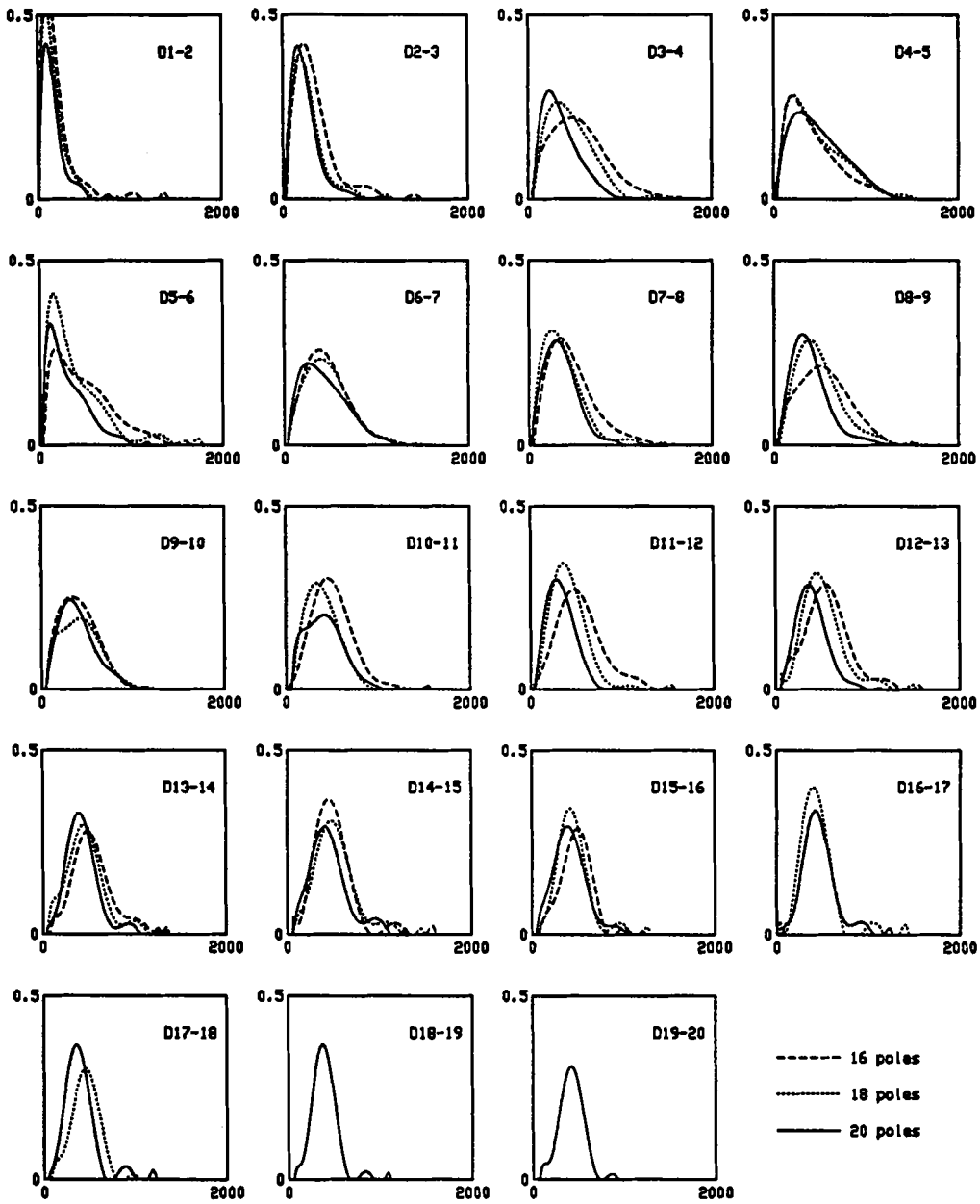


Fig. 3.2 Adjacent LSF spectral distance densities for 16, 18 and 20 poles.

not *spill over* the upper bound of 8000 Hz. This could happen if a small number of bits is used to quantize the distances between these LSF's. Also, even and odd LSF's must be given similar resolutions. The formant structure in speech is determined by the location and relative spacing of odd and even LSF pairs. Thus, proper resolu-

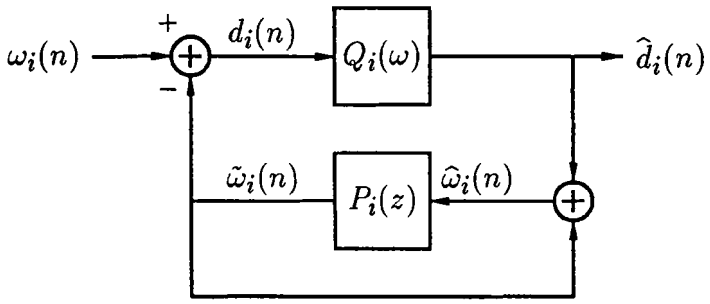
tions must be provided to minimize the formant shifts and the formant bandwidth distortion.

3.2.3 Time differential non–uniform quantization (TDNUQ)

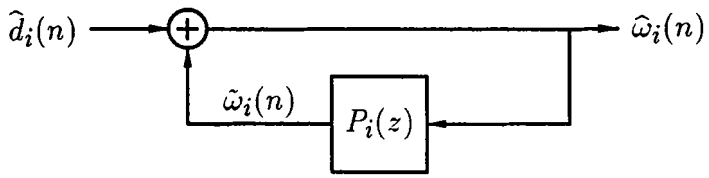
Another LSF quantization scheme, suggested by Crosmer and Barnwell [16], uses both time and frequency differences between LSF's. The use of LSF time spectral distances is motivated by a strong LSF frame to frame correlation in steady speech segments. A DPCM (Differential Pulse Code Modulation) coder can efficiently exploit this by quantizing the difference between LSF's and their predicted values (Figure 3.3). The resulting time spectral distances have smaller dynamic ranges, thereby increasing the quality of the quantizer.

However, only odd LSF's are coded with DPCM. Since the formant structures in speech are defined by close odd and even LSF pairs, and since the formants do not all vary in the same direction from frame to frame, especially at low frequencies, the use of independent DPCM coders for each LSF could to lead badly ordered quantized LSF's. Crosmer and Barnwell used DPCM coding for odd LSF's, and then quantized the relative spectral distance between even LSF's and their odd neighbors.

In the TDNUQ method suggested here, fixed predictors (i.e. $P_i(z) = 1$ are used in the DPCM coders. The non–uniform quantizers Q_i are based on the odd LSF's time spectral densities shown in Figure 3.4. Each even LSF ω_i is adaptively encoded from an M–level uniform quantizer based on the range $[\widehat{\omega}_{i-1}, \widehat{\omega}_{i+1}]$. The use of a uniform quantizer assumes accurate DPCM coding of the odd LSF's and, if good enough, all quantized LSF's tend to be properly ordered.



(a) coder



(b) decoder

Fig. 3.3 DPCM coding/decoding for LSF quantization.

3.3 LSF Quantization Tests

The three LSF quantization structures presented in Section 3.2 (NUQ, DNUQ, TDNUQ) are now tested. The goal of these tests is to establish a relationship between the performance of each proposed quantization scheme and the number of bits required per frame. The tests are described below and the results follow.

3.3.1 Description

A total of 24 sentences, spoken by 2 males and 2 females are used. For each LSF quantization method, the number of bits per frame is set to 28, 40, 50, 60 and 75. In

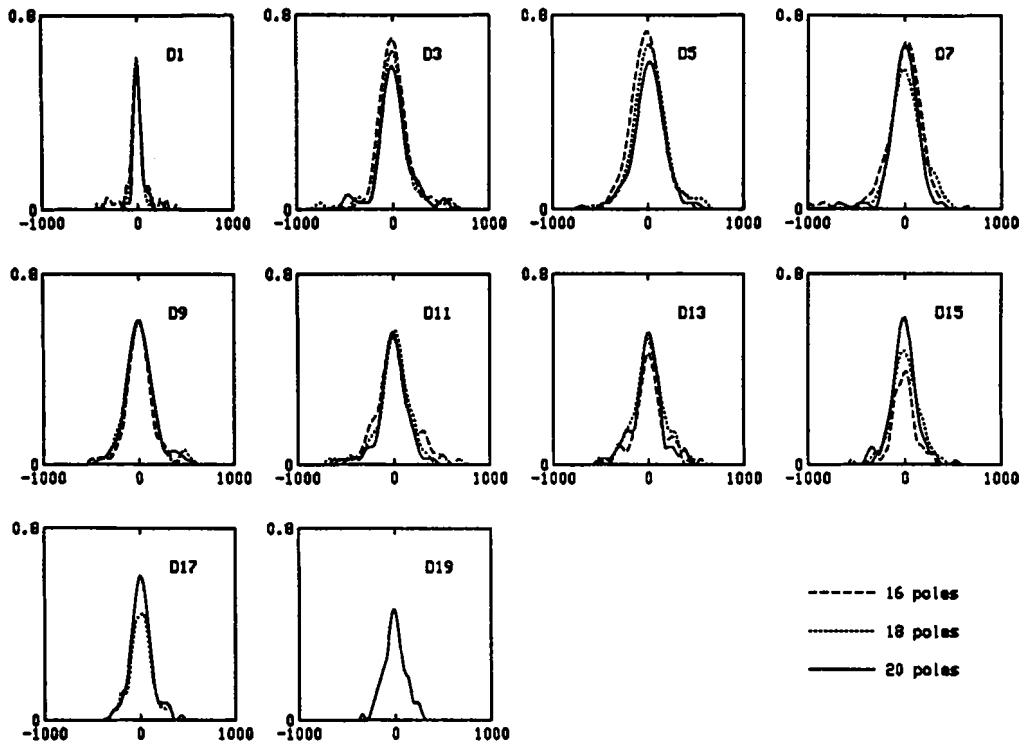


Fig. 3.4 Time spectral densities of odd LSF for 16, 18 and 20 poles.

each case, the distribution of the bits between the low band (0–4 kHz) and the high band (4–8 kHz) is varied from 50/50% to 70/30%.

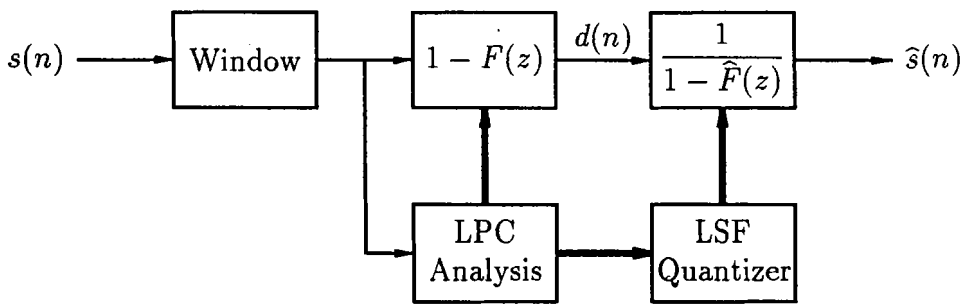


Fig. 3.5 LSF quantization test coder structure.

The coder structure used for the tests is shown in Figure 3.5. A 25 ms Hamming

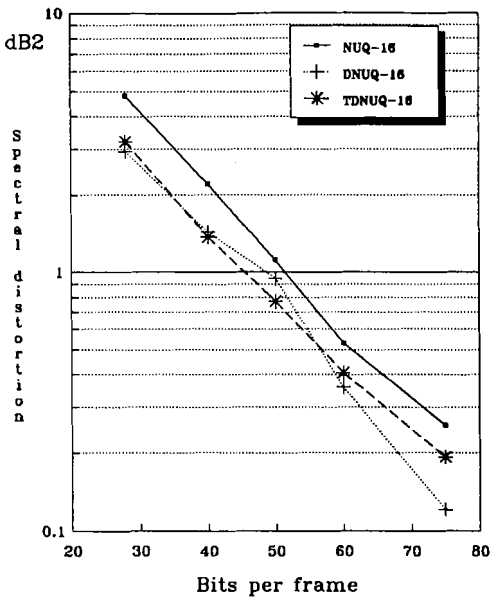
window is applied to the input signal $s(n)$. For each 20 ms frame of speech, the LSF's are computed and the windowed signal is inverse filtered by the error filter $1 - F(z)$ to produce the formant residual signal $d(n)$. This residual is then directly fed back to excite the all-pole synthesis filter $\widehat{H}(z) = 1/(1-\widehat{F}(z))$ defined by the quantized LSF's. This arrangement is useful in determining the performance of the LSF quantization scheme used. Two performance measures are used: *Segmental Signal to Noise Ratio* (SegSNR), and the *average Spectral Distortion* (SD). The SegSNR is an SNR measure calculated in dB and averaged over frames of 20 ms duration. The SD measure is calculated in dB^2, and is defined as:

$$SD = \frac{1}{N_f} \sum_{n=1}^{N_f} \left[\frac{1}{\pi} \int_0^{\pi} (\log S_n(\omega) - \log \widehat{S}_n(\omega))^2 d\omega \right] \tag{3.8}$$

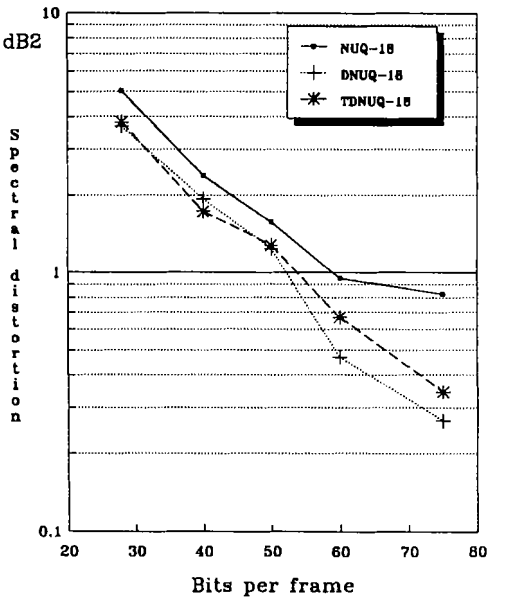
where $S_n(\omega)$ and $\widehat{S}_n(\omega)$ are respectively the unquantized and quantized speech spectra for the n^{th} frame, and N_f is the total number of frames. In general, the quantization effects become negligible when the spectral distortion falls under the difference limen value of 1 dB^2 [16]. Although this is a good indication of performance, it remains an averaged measure. As such, it may not always reflect the presence of larger distortion levels in some isolated frames for which the perceived quality could be reduced.

3.3.2 Results

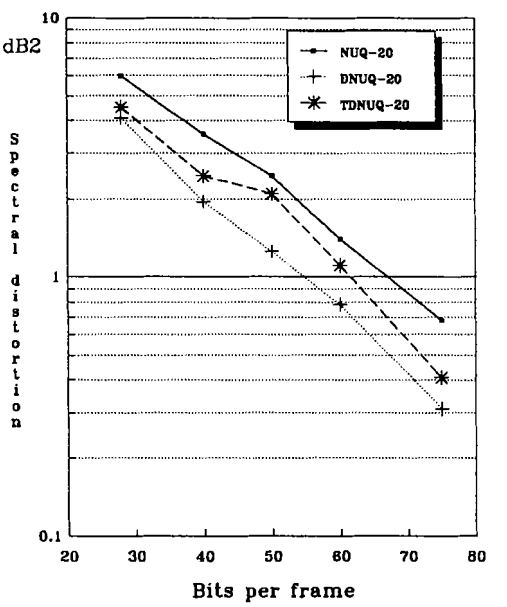
The spectral distortion and segmental SNR results for each method are comparatively plotted in Figures 3.6 and 3.7 for different numbers of poles . One immediate result is that both the SD and SegSNR measures deteriorate as the number of poles increases. This is expected, as the effective number of bits per parameter decreases.



(a) 16 poles



(b) 18 poles



(c) 20 poles

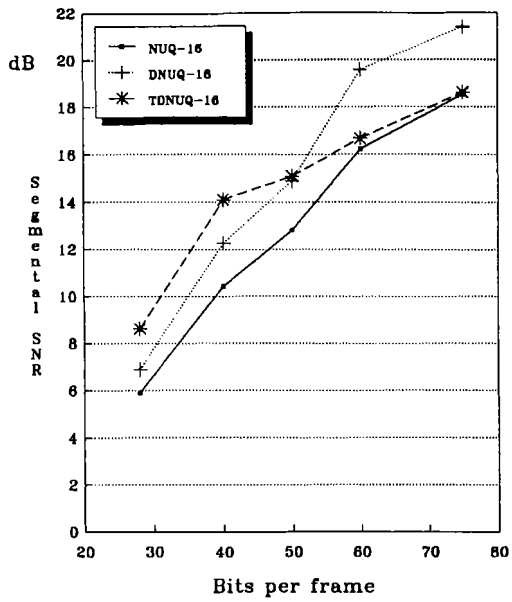
Fig. 3.6 Comparative spectral distortion measures of NUQ, DNUQ and TDNUQ for 16, 18 and poles.

The SD graphs of Figure 3.6(a), (b) and (c) show that the NUQ method performs consistently worse than the other two, especially at low rates (i.e. at less than 40 bits per frame). This is a direct consequence of the limitations of NUQ, as this method cannot accurately quantize independent LSF's with wide and overlapping dynamic ranges.

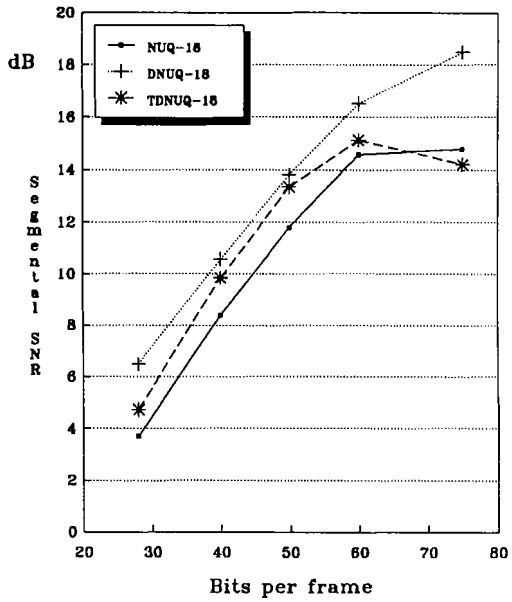
The DNUQ and TDNUQ methods, however, performed similarly for 16 and 18 poles systems, while DNUQ has the advantage at 20 poles. Note that for a SD of 1 dB^2, the minimum number of bits per frame is around 50 for 16 poles, 55 for 18 poles and 60 for 20 poles.

The SegSNR graphs of Figure 3.7(a), (b) and (c) indicate that the DNUQ method outperforms the other two, except TDNUQ-16 at rates less than 50 bits per frame. Perhaps this is due to the uniform quantization scheme used in TDNUQ to quantize the even LSF's. As tested, the TDNUQ method allocates the bits equally between the odd and even LSF's. This is beneficial at low rates (i.e. less than 40 bits per frame), as the DPCM quantizers need few bits to properly code the odd LSF's. But as the rate increases, the gained resolution for coding the odd LSF's is not as important as that of the even LSF's. This could explain the changing TDNUQ performance at different rates.

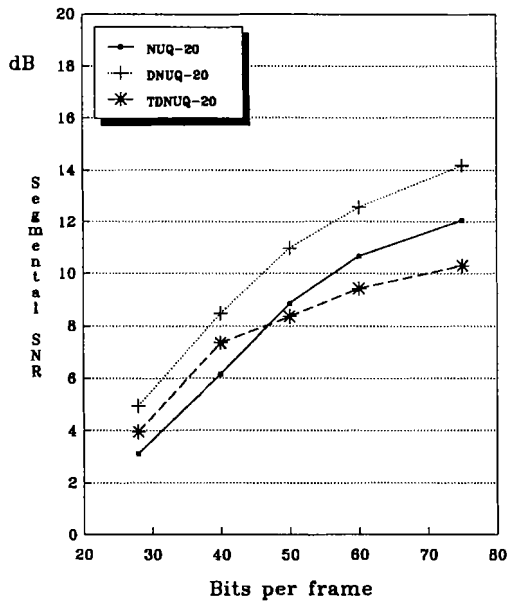
To verify this, the TDNUQ method was modified in 3 ways. First, more bits were given to the even than to the odd LSF's. Second, the uniform quantizers of the even LSF's were replaced by the non–uniform quantizers developed for DNUQ: the odd LSF's were first coded with DPCM, then the even LSF's were coded with respect to their preceding odd LSF's with DNUQ. Finally, all LSF's were coded with the



(a) 16 poles



(b) 18 poles



(c) 20 poles

Fig. 3.7 Comparative SegSNR measures of NUQ, DNUQ and TDNUQ for 16, 18 and poles.

DPCM scheme. In all 3 cases, the results were better than with standard TDNUQ, except at very low rates, where more crossovers occurred. For rates greater than 40 bits per frame, the second modification was the best, with improvements between 1 and 2 dB for SegSNR and between 0.2 and 0.1 dB^2 for SD. However, none of the 3 modifications yielded better overall results than the standard DNUQ.

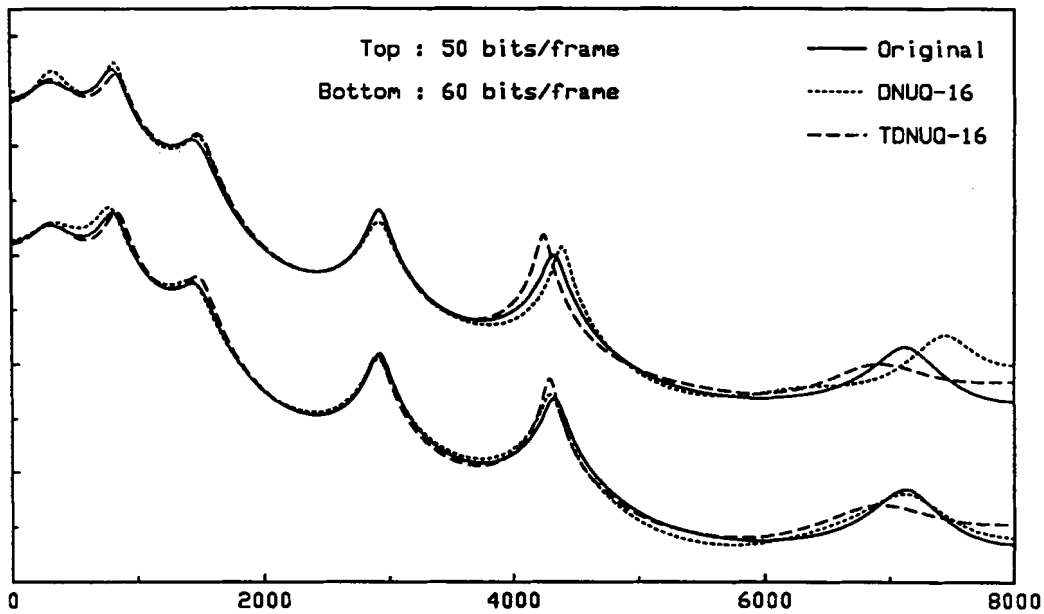
Among the best performing methods, it is interesting to note that the SD and SegSNR measures vary differently with respect to the bit allocation between the low and high frequency bands. In particular, the SegSNR measures are generally better when more emphasis is put on the low band, e.g. 67/33% for DNUQ. The SD however, is slightly better when the distribution lies around 60/40%.

For all methods, better SegSNR and worse SD figures were obtained when the last LSF is given 0 bits, i.e. when it is left to its mean value. This is expected for the SegSNR, as the bits normally used for the last LSF are now applied to the lower frequency components which tend to carry more energy. However, this also tends to show that the SD measure, although suitable for narrowband systems in its current form, may not be appropriate for wideband systems. The SD measure reflects the distortion level between the original and the coded spectral envelopes, but does not take into account the frequency range. Therefore, a large distortion level induced at the last LSF has a negative impact on the overall SD measure when it is, in fact, hardly noticeable in listening tests. Perhaps a weighting function should be applied to the current SD measure to reflect the lower perceptual impact of the higher frequencies found in wideband signals.

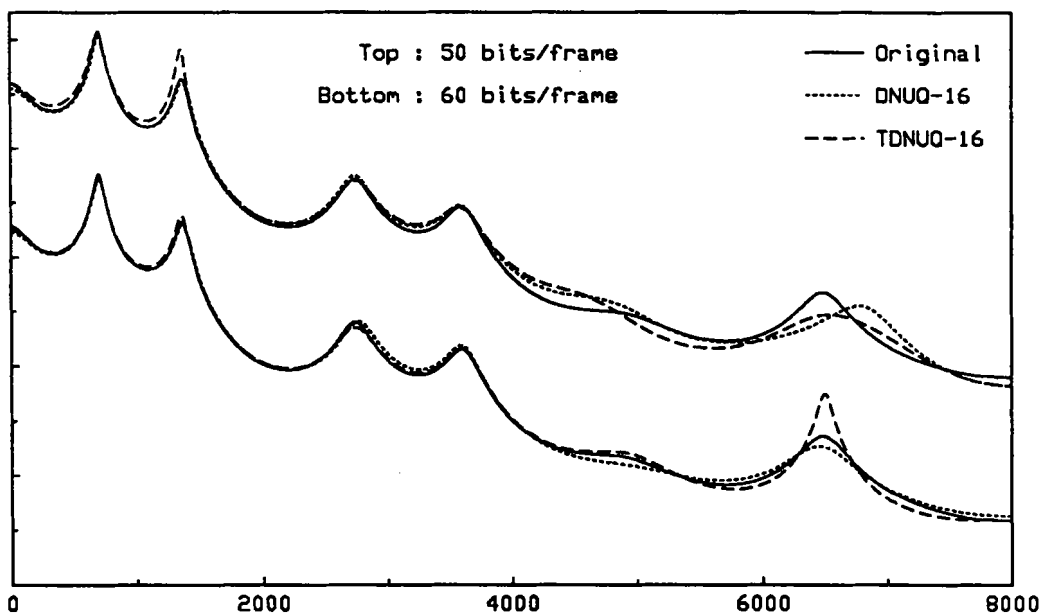
Comparative spectral envelope plots are shown in Figures 3.8 to 3.10 for a male

and a female speaker for rates of 50 and 60 bits per frame. First, the performance of each method gets worse when 20 poles are used. The spectral envelope in the higher band is not well modeled, especially for the TDNUQ-20 method. This is a direct consequence of the reduced number of bits per pole. The difference is not so clear between 16 and 18 poles systems, although in either case, 60 bits/frame yields better results than 50 bits/frame. The DNUQ-18 and TDNUQ-18 methods both start to show degradations around 3000 Hz. This only happens around 4200 Hz for DNUQ-16 and TDNUQ-16. Finally, for the female segment, the DNUQ approach shows undesirable distortion while modeling the first 3 formants. This is a case where the low formants are very close to each other, and the non–uniform quantizers have difficulty coding the small distances between adjacent LSF's.

These results indicate that at operating rates of 50 bits/frame, no more than 16 poles should be used. Although the methods presented in this chapter are not the most efficient in terms of bit rate, they still help determine a basic rate for coding the wideband spectral envelope. This estimate will be used in the simulations of Chapter 5 to study wideband coders operating at 16 kbits/sec.

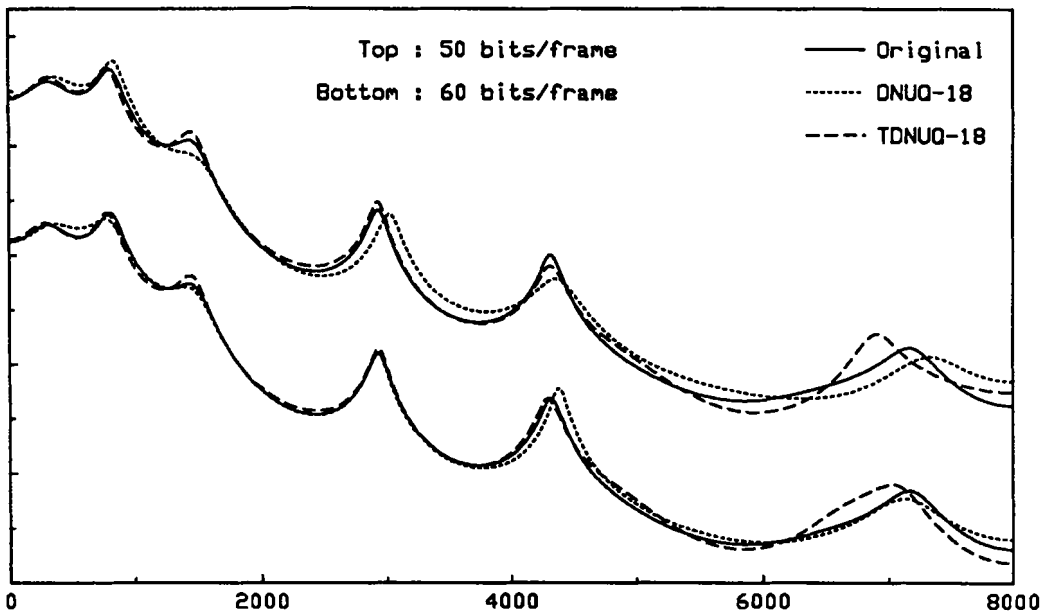


(a) female speech segment

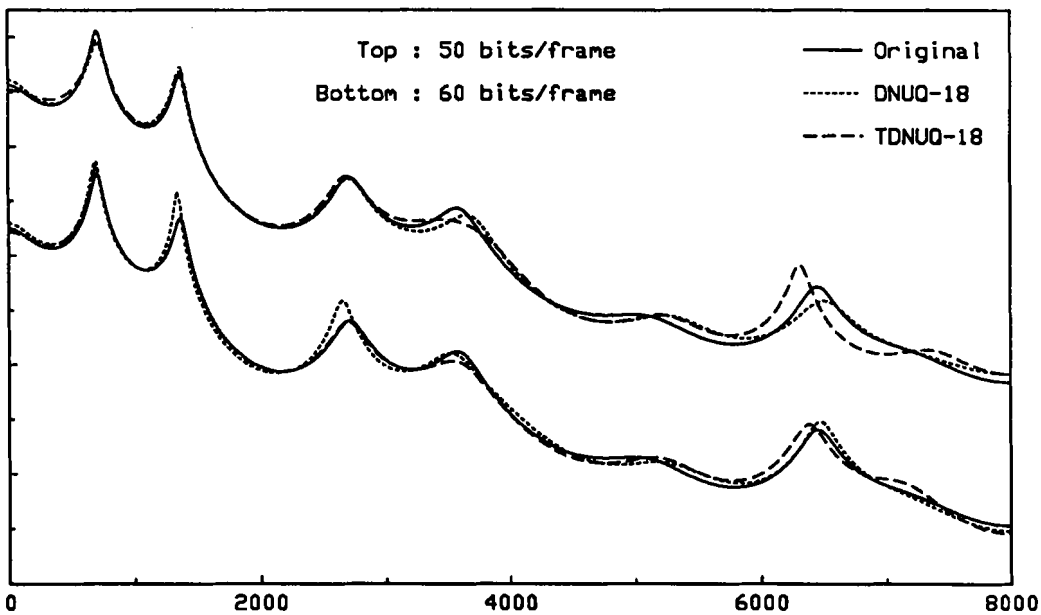


(b) male speech segment

Fig. 3.8 DNUQ and TDNUQ spectral envelopes for 16 poles.

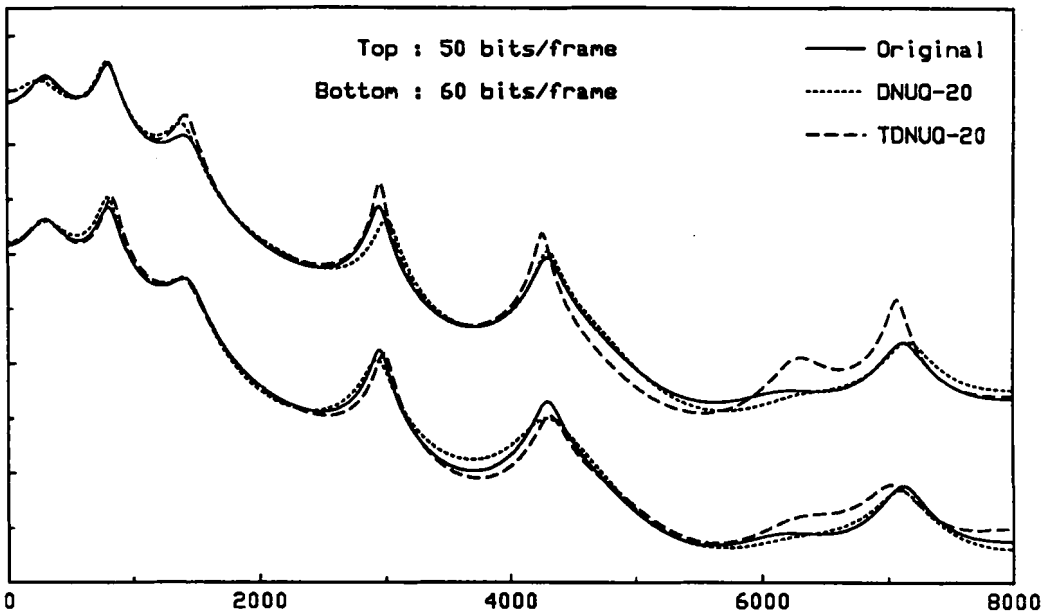


(a) female speech segment

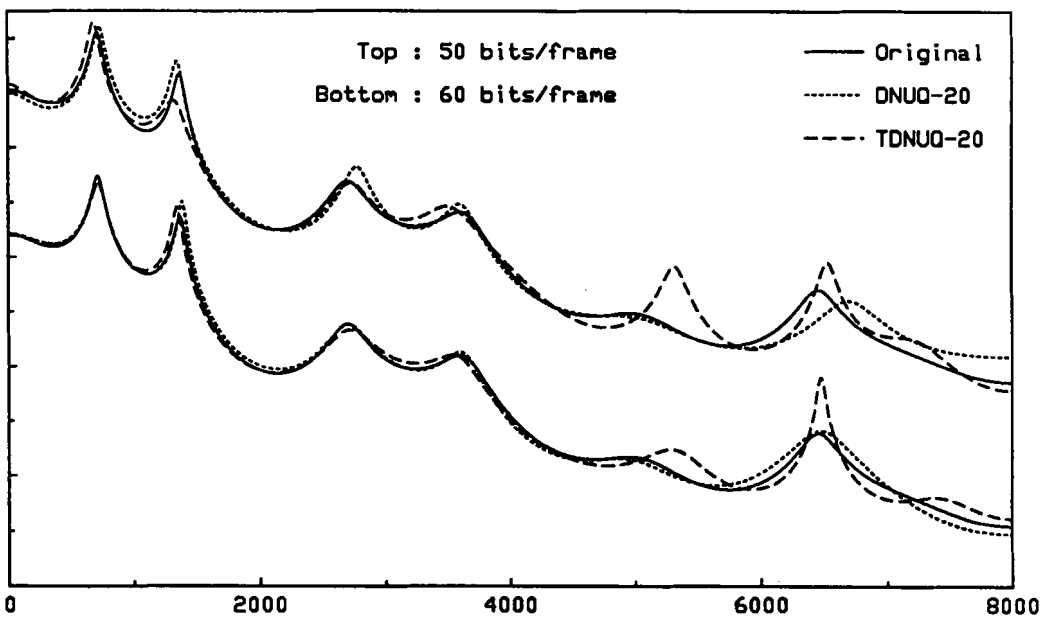


(b) male speech segment

Fig. 3.9 DNUQ and TDNUQ spectral envelopes for 18 poles.



(a) female speech segment



(b) male speech segment

Fig. 3.10 DNUQ and TDNUQ spectral envelopes for 20 poles.

Chapter 4

Enhanced Wideband RELP Coding

In this chapter, the basic RELP model is revisited and applied in a wideband context. First, it is enhanced by the addition of a pitch prediction stage, then transformed into an analysis–by–synthesis structure by re-optimizing the pitch prediction parameters. Although these approaches constitute a fresh look at RELP, the intent is not to produce a functional wideband RELP coder. Rather, it is to demonstrate that the high frequency speech components can be reproduced, with acceptable quality, using sub-optimal excitation waveforms. This helps lay down the basis for the wideband implementation of a CELP coder. To this effect, parameter quantization is left aside, except for the last section, where residual coding is investigated, and migration to CELP justified.

4.1 Addition of a Pitch Prediction Stage

The introduction of long term prediction reduces the variance of the residual signal and removes most of its harmonic structure. Consider the RELP model shown in Figure 4.1.

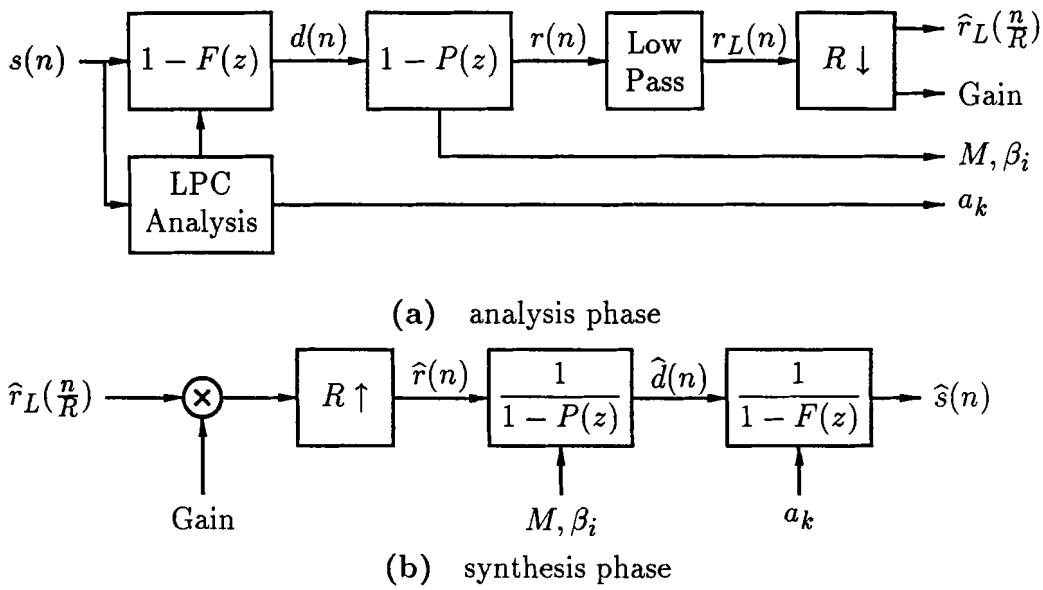


Fig. 4.1 RELP coder with pitch prediction.

As described in Section 2.2, the quality of the reproduced speech $\hat{s}(n)$ is directly affected by the High Frequency Regeneration stage. The HFR stage in this model is reduced to simple upsampling by a factor R. The excitation signal $\hat{r}(n)$ at the receiver is then made up of copies of the baseband $\hat{r}_L(n)$ spectrally folded through the whole band.

The addition of the pitch prediction stage reduces some of the degradations encountered in the basic RELP model with spectral folding. In particular, $\hat{s}(n)$ does not sound as metallic. The pitch structure is re-introduced after upsampling and it does not suffer from the discontinuity problems originally found at the spectral folding junctions. Also, most of the clicks and pops are replaced by a uniform degradation, reminiscent of the noise introduced by low-level quantization. It is important to note though, that these effects are most noticeable when a small baseband is kept (i.e. 1 kHz). They generally disappear when 4 kHz of baseband residual is preserved.

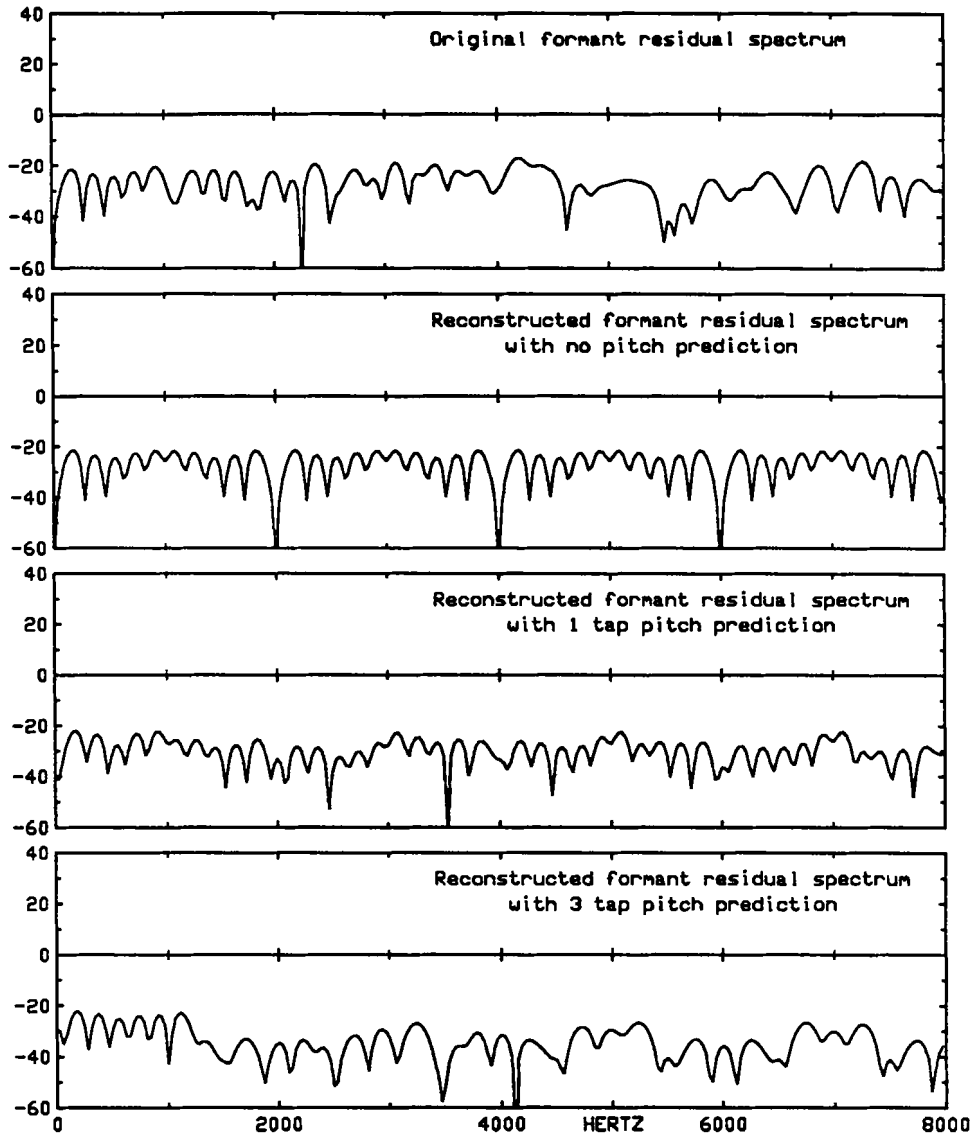


Fig. 4.2 Regenerated formant residual spectrums for RELP with 0, 1 and 3 tap pitch prediction.

Consider the plots of Figures 4.2 and 4.3, obtained when 1 kHz of baseband is transmitted and pitch prediction orders of 0, 1 and 3 are used. Figure 4.2 shows the original formant residual signal $d(n)$ in the top trace, whereas the other traces contain the regenerated formant residuals $\widehat{d}(n)$ for 0, 1 and 3 tap pitch prediction respectively. Figure 4.3 shows the original formant speech signal $s(n)$ in the top trace, whereas the

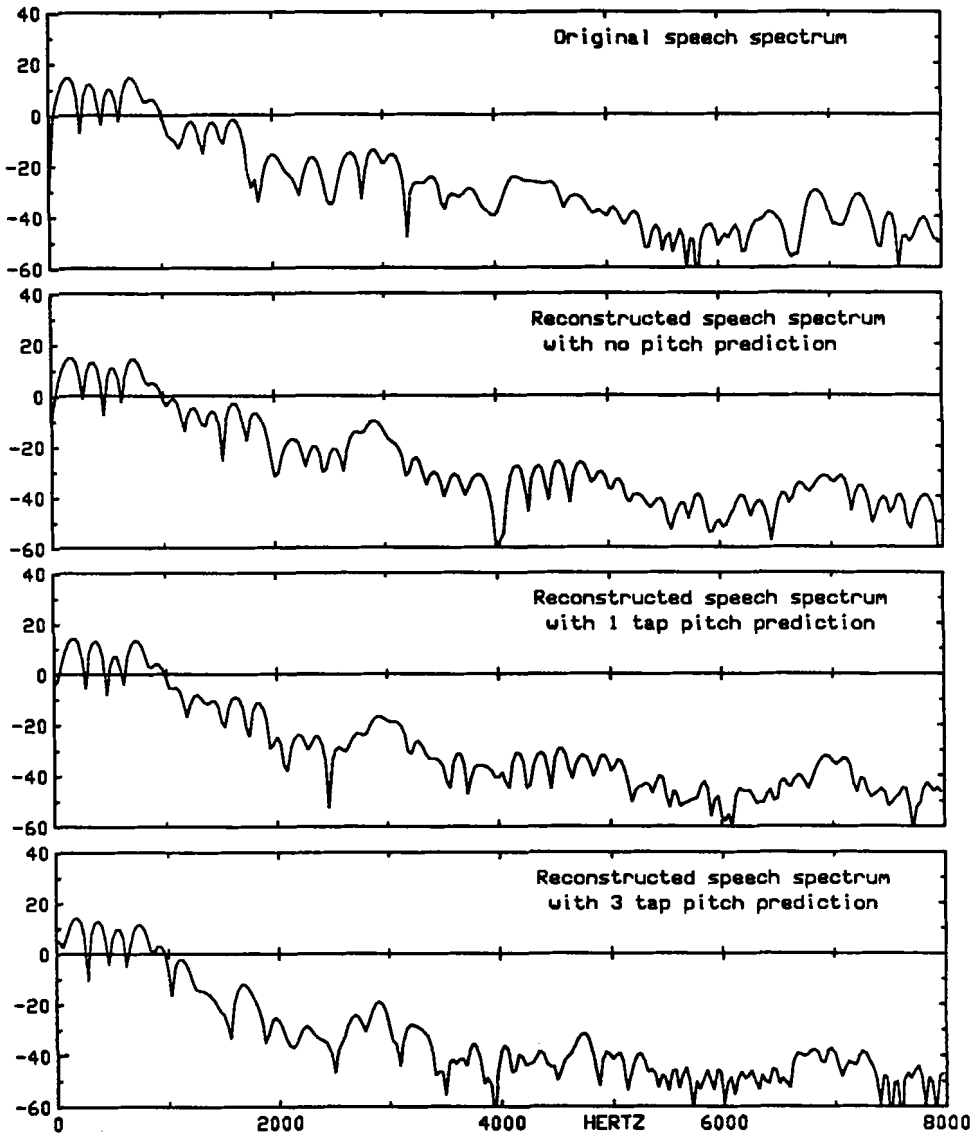


Fig. 4.3 Reconstructed speech spectrums for RELP with 0, 1 and 3 tap pitch prediction.

other traces contain the regenerated speech $\hat{s}(n)$ for 0, 1 and 3 tap pitch prediction respectively.

The effects of HFR are visible in both figures when no pitch prediction is used. The pitch structure shows discontinuities, especially at 2, 4 and 6 kHz boundaries. When a 1 tap pitch predictor is used, most of the discontinuities disappear. Some

remain since the pitch residual signal $\hat{r}(n)$ still contains a small, but noticeable fine line structure. The effects of interpolation are therefore still visible. When a 3 tap pitch prediction filter is used, these discontinuities are essentially gone.

4.2 Optimization of the Pitch Prediction Stage

Although it seems that the addition of pitch prediction can simplify the HFR scheme and yield better reconstructed speech, it also introduces a few problems. In particular, stability, as discussed in Section 2.1.2, becomes an issue. In order to ensure that the pitch synthesis filter is stable, the pitch coefficients β_i must be tested as prescribed by Ramachandran [5]. Therefore, the stabilized coefficients are no longer optimal, and may not remove as much of the pitch structure as they normally would.

Another problem affecting the quality of the reproduced speech is the modification of the relation between the prediction coefficients and the residual. In a linear system, given a set of formant and pitch parameters, the original input signal can be exactly reproduced provided the residual signal is not modified. From a practical point of view, this is never possible, since the residual signal must be quantized and, in this case, low-pass filtered and decimated. The excitation signal $\hat{r}(n)$ appearing before the pitch and formant synthesis filters is no longer optimal with respect to the original pitch and formant synthesis coefficients.

This situation arises since the residual signal $\hat{r}(n)$ is determined *after* the prediction coefficients. The problem is now inverted and the question is: *"Given the excitation $\hat{r}(n)$, can optimal pitch and formant parameters yielding a reconstructed*

speech signal $\hat{s}(n)$ *identical to the original signal* $s(n)$ *be found?"*. The answer is no. However, it is possible to modify the parameters to obtain the best possible reconstructed speech.

In the case of formant parameters, the re-optimization process becomes an exercise in solving high-order non-linear equations. These equations are induced by the low–delay feedback loop of the formant synthesis filter. Although there exist iterative methods to solve this kind of problem, a globally optimum solution cannot be ensured. For simplicity, the formant parameters are therefore not subjected to re-optimization. The pitch parameters however, can be re-optimized. The following two sub-sections describe the procedures necessary for obtaining the optimal set of pitch parameters. In the first sub-section, the residual excitation consists of a single signal $\hat{r}(n)$ occupying the full bandwidth. In the second, the excitation source is made up of two separate excitation residuals $\hat{r}_L(n)$ and $\hat{r}_H(n)$, respectively occupying the low and high frequency bands.

4.2.1 Full–band optimization

Consider the model shown in Figure 4.4. Let the formant coefficients a_k be as defined in the analysis phase of Figure 4.1. However, let the excitation residual $\hat{r}(n)$ be the unscaled, interpolated version of $\hat{r}_L(n/R)$.

As in the basic CELP model presented in Section 2.3, γ represents the bandwidth expansion factor. In contrast with the basic CELP model, the error weighting filter $W(z)$ has been incorporated within each branch. Also, the excitation waveform $\hat{r}(n)$

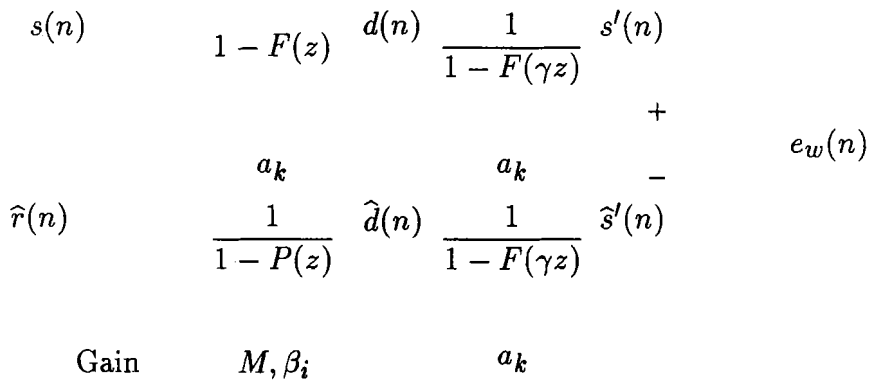


Fig. 4.4 Full–band RELP pitch optimization.

is already known from the analysis stage. For each pitch sub-frame, the pitch coefficients β_i, the pitch lag M and the gain G that minimize the energy of the weighted error $e_w(n)$, given $\hat{r}(n)$, must then be found. At the receiver, the speech is reconstructed exactly as shown in Figure 4.1(b). The procedure for finding the optimal pitch parameters is similar to the one derived in [17], and is described below.

Let the weighted error signal $e_w(n)$ be defined as:

$$e_w(n) = s'(n) - \hat{s}'(n), \tag{4.1}$$

where the bandwidth expanded original and reconstructed speech signals $s'(n)$ and $\hat{s}'(n)$ are obtained, by convolution, as follows:

$$s'(n) = \sum_{k=-\infty}^{\infty} d(k)h'(n-k), \tag{4.2}$$

$$\hat{s}'(n) = \sum_{k=-\infty}^{\infty} \hat{d}(k)h'(n-k). \tag{4.3}$$

The impulse response of the bandwidth expanded formant synthesis filter $h'(n)$ is derived by geometrically scaling the original formant predictor coefficients a_k by

the bandwidth expansion factor γ. Equation 2.3 is thus modified as follows:

$$\begin{aligned} H'(z) &= H(\gamma z) \\ &= \frac{1}{1 - F(\gamma z)} \\ &= \frac{1}{1 - \sum_{k=1}^{P} a_k (\gamma z)^{-k}} \\ &= \frac{1}{1 - \sum_{k=1}^{P} a'_k z^{-k}} \\ &= \frac{1}{1 - F'(z)}. \end{aligned} \tag{4.4}$$

When γ takes on values larger than 1 (e.g. $\gamma = 1/0.75$), this operation amounts to enlarging the unit circle to a radius of length γ. Conversely, this can be seen as radially shifting all the poles inward. The filter stability is preserved since the new poles remain within the unit circle. The shifted poles yield wider spectral valleys than those found with the original poles. This concentrates the coding distortion in the spectral valleys where it is better masked by the surrounding spectral peaks. The impulse response $h'(n)$ is also time-varying. However, since the minimization procedure is done at the pitch sub-frame level, $h'(n)$ is known and held constant for the duration of the sub-frame.

In Equation 4.2, both the formant residual $d(n)$ and the impulse response $h'(n)$ are fully known for all values of k. Moreover, the summation limits can be changed to 0 and $N-1$, the sub-frame length, provided that the contribution of past sub-frame excitation samples (i.e. $k < 0$) are preserved as initial conditions for the current sub-frame. This is achieved by saving the IIR filter internal memory from one sub-frame to the next.

In equation 4.3, $\hat{s}'(n)$ can be broken into its anti-causal and causal parts:

$$\hat{s}'(n) = \sum_{k=-\infty}^{-1} \hat{d}(k)h'(n-k) + \sum_{k=0}^{\infty} \hat{d}(k)h'(n-k). \tag{4.5}$$

The first summation term is the anti-causal response, or the output of the bandwidth expanded formant synthesis filter due to past excitation values. This is obtained by letting the IIR filter *free-wheel* with its internal values while feeding it a null excitation. In other words, it is the zero-input response and it accounts for initial conditions at sub-frame boundaries. This term is fully known for the duration of the sub-frame.

The causal term however, depends on the regenerated formant residual signal $\hat{d}(n)$. This signal is the output of the pitch synthesis filter, and is expressed as:

$$\hat{d}(n) = G\hat{r}(n) + \sum_{i=1}^{N_p} \beta_i \hat{d}(n-M-i), \tag{4.6}$$

where G is the gain factor, N_p is the number of pitch coefficients and M is the pitch lag. Although this is an IIR filter structure, non-linear recursions can be prevented by forcing the pitch lag M to be greater than the pitch sub-frame length N. This effectively *cuts* the feedback path of the pitch synthesis filter. Then, $\hat{d}(n)$ can be viewed as a linear combination of the fully known waveforms $\hat{r}(n)$ and $\hat{d}(n-M-i)$ for all values of i. Substituting Equation 4.6 into the causal term of 4.5 yields:

$$\begin{aligned} \hat{s}'(n) = \sum_{k=-\infty}^{-1} \hat{d}(k)h'(n-k) + G\sum_{k=0}^{\infty} \hat{r}(k)h'(n-k) \\ + \sum_{i=1}^{N_p} \beta_i \sum_{k=0}^{\infty} \hat{d}(k-M-i)h'(n-k). \end{aligned} \tag{4.7}$$

The above equation can be expressed more simply as:

$$\hat{s}'(n) = \sum_{k=-\infty}^{-1} \hat{d}(k)h'(n-k) + Gx(n) + \sum_{i=1}^{N_p} \beta_i\, y_i(n), \tag{4.8}$$

where the following substitution are made:

$$\begin{aligned} x(n) &= \sum_{k=0}^{\infty} \hat{r}(k)h'(n-k), \\ &= \sum_{k=0}^{N-1} \hat{r}(k)h'(n-k), \end{aligned} \tag{4.9}$$

and

$$\begin{aligned} y_i(n) &= \sum_{k=0}^{\infty} \hat{d}(k-M-i)h'(n-k), \\ &= \sum_{k=0}^{N-1} \hat{d}(k-M-i)h'(n-k). \end{aligned} \tag{4.10}$$

In the above two equations, the upper limits of summation of $x(n)$ and $y_i(n)$ have been changed to $N-1$. This is valid since the impulse response $h'(n)$ is causal and the minimization is done over the finite sub-frame interval of N samples. Therefore, Equation 4.1 can be rewritten as:

$$\begin{aligned} e_w(n) &= s'(n) - \sum_{k=-\infty}^{-1} \hat{d}(k)h'(n-k) - Gx(n) - \sum_{i=1}^{N_p} \beta_i y_i(n), \\ &= s^*(n) - Gx(n) - \sum_{i=1}^{N_p} \beta_i y_i(n), \end{aligned} \tag{4.11}$$

where $s^*(n)$ contains all the terms not subjected to optimization.

The minimization is done in the mean-square sense. Let the energy of the weighted error signal $e_w(n)$ in the pitch sub-frame be:

$$\xi = \sum_{n=0}^{N-1} e_w^2(n). \tag{4.12}$$

Differentiating the above equation with respect to the gain and the pitch coefficients and setting it equal to 0 yields, for any given pitch lag value M, a set of linear

equations. Using the chain rule of differentiation, these are obtained as follows:

$$\begin{aligned}\frac{\delta\xi}{\delta G} &= \sum_{n=0}^{N-1}\left[2\left[s^*(n) - Gx(n) - \sum_{i=1}^{N_p}\beta_i y_i(n)\right]\left[-x(n)\right]\right] \\ &= -2\sum_{n=0}^{N-1} s^*(n)x(n) + 2G\sum_{n=0}^{N-1} x(n)^2 \\ &\quad + 2\sum_{n=0}^{N-1}\sum_{i=1}^{N_p}\beta_i y_i(n)x(n) \\ &= 0,\end{aligned} \tag{4.13}$$

and, for $j = 1 \ldots N_p$,

$$\begin{aligned}\frac{\delta\xi}{\delta\beta_j} &= \sum_{n=0}^{N-1}\left[2\left[s^*(n) - Gx(n) - \sum_{i=1}^{N_p}\beta_i y_i(n)\right]\left[-y_j(n)\right]\right] \\ &= -2\sum_{n=0}^{N-1} s^*(n)y_j(n) + 2G\sum_{n=0}^{N-1} x(n)y_j(n) \\ &\quad + 2\sum_{n=0}^{N-1}\sum_{i=1}^{N_p}\beta_i y_i(n)y_j(n) \\ &= 0.\end{aligned} \tag{4.14}$$

Equations 4.13 and 4.14 can be simplified, and expressed in matrix form as $\Phi\mathbf{v} = \mathbf{b}$, where the matrix Φ and the vectors $\mathbf{v}$ and $\mathbf{b}$ are defined as †:

$$\Phi = \begin{pmatrix} \langle x(n)^2\rangle & \langle y_1(n)x(n)\rangle & \cdots & \langle y_{N_p}(n)x(n)\rangle \\ \langle x(n)y_1(n)\rangle & \langle y_1(n)^2\rangle & \cdots & \langle y_{N_p}(n)y_1(n)\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle x(n)y_{N_p}(n)\rangle & \langle y_1(n)y_{N_p}(n)\rangle & \cdots & \langle y_{N_p}(n)^2\rangle \end{pmatrix}, \tag{4.15}$$

$$\mathbf{v} = \begin{pmatrix} G \\ \beta_1 \\ \vdots \\ \beta_{N_p} \end{pmatrix}, \tag{4.16}$$

$$\mathbf{b} = \begin{pmatrix} \langle s^*(n)x(n)\rangle \\ \langle s^*(n)y_1(n)\rangle \\ \vdots \\ \langle s^*(n)y_{N_p}(n)\rangle \end{pmatrix}. \tag{4.17}$$

† For clarity, all summations symbols are left out. Thus, $\langle x(n)\rangle$ refers to $\sum_{n=0}^{N-1} x(n)$

The covariance matrix Φ can be written more simply as:

$$\Phi = \langle \mathbf{q}\mathbf{q}^{\mathrm{T}} \rangle, \tag{4.18}$$

where

$$\mathbf{q}^{\mathrm{T}} = \left(x(n), y_1(n), \ldots, y_{N_p}(n) \right). \tag{4.19}$$

The solution to this system is found using the Cholesky factorization algorithm. In certain situations however, Φ may be ill-conditioned and the solution is then unreliable. This is the case when the residual excitation $\hat{r}(n)$ is null, thereby canceling the first entry in $\mathbf{q}^{\mathrm{T}}$ (see Eq. 4.9). A similar situation arises when the past regenerated formant residuals $\hat{d}(n - M - i)$ are close to, or equal to zero, in which case some of the $y_i(n)$ may also be close to, or equal to zero (see Eq. 4.10). Inversion problems are avoided by monitoring the diagonal entries in Φ. For each diagonal entry close or equal to 0, the corresponding variable in the solution vector $\mathbf{v}$ is set to 0 and eliminated from the system of equations. This reduces the order of the system by 1. For example, suppose that $\phi_{22} = 0$. Then, β_1 is set to 0, and the reduced system $\Phi'\mathbf{v}' = \mathbf{b}'$ is defined as:

$$\Phi' = \langle \mathbf{q}'\mathbf{q}'^{\mathrm{T}} \rangle, \tag{4.20}$$

where

$$\mathbf{q}'^{\mathrm{T}} = \left(x(n), y_2(n), \ldots, y_{N_p}(n) \right). \tag{4.21}$$

$$\mathbf{v}' = \begin{pmatrix} G \\ \beta_2 \\ \vdots \\ \beta_{N_p} \end{pmatrix}, \tag{4.22}$$

$$\mathbf{b}' = \begin{pmatrix} \langle s^*(n)x(n)\rangle \\ \langle s^*(n)y_2(n)\rangle \\ \vdots \\ \langle s^*(n)y_{N_p}(n)\rangle \end{pmatrix}. \tag{4.23}$$

This reduced system is then solved using the Cholesky factorization algorithm. Finally, the solution vector $\mathbf{v}$ depends on the pitch lag M since $y_i(n)$ is a function of M (see Eq. 4.10). The overall optimal solution is thus obtained by forming, then solving the matrix system for each possible lag values within the allowed pre-defined lag range.

4.2.2 Split–band optimization

The optimization procedure in the previous section assumes that the sub-optimal excitation $\hat{r}(n)$ is uniformally degraded in frequency. This is not always the case, especially when the decimated residual is quantized well before being transmitted. Thus, the baseband portion of the excitation residual $\hat{r}(n)$ matches the original baseband within the restrictions imposed by quantization. The upper band portion of $\hat{r}(n)$ however, does not match its original counterpart. The calculated optimal pitch parameters therefore compensate for the sub-optimality of the upper band, but can introduce unnecessary distortion in the resulting baseband speech.

The structure presented in Figure 4.5 separately optimizes the pitch parameters for each band. In part (a), the decimated baseband residual $\hat{r}_L(n/R)$ obtained at the analysis stage (see Fig. 4.1(a)) is first upsampled, then separated by a pair of complementary low-pass and high-pass filters. The parameter optimization is carried out in part (b). At the receiver, the residual $\hat{r}_L(n/R)$ is split as shown in part (a), with

$\hat{r}_L(n)$ and $\hat{r}_H(n)$ respectively exciting the optimal low and high band pitch synthesis filters. The speech is reconstructed by exciting the formant synthesis filter with the sum of the regenerated low and high band formant residuals $\hat{d}_L(n)$ and $\hat{d}_H(n)$, as shown in part (c).

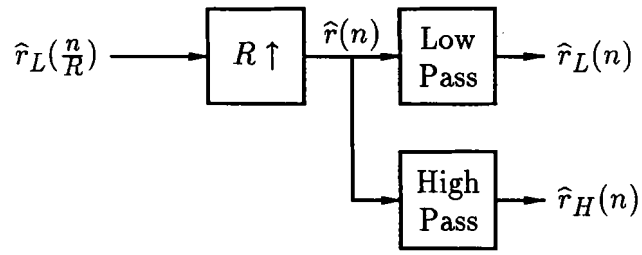
The procedure for finding the optimal split-band parameters is an extension of the one presented in the Section 4.2.1. Here also, the goal is to find the set of pitch parameters that minimize the energy of the weighted error $e_w(n)$. However, the regenerated formant residual $\hat{d}(n)$ is now expressed as:

$$\begin{aligned}\hat{d}(n) = G_L\hat{r}_L(n) + \sum_{i=1}^{N_{p_L}} \beta_{L,i}\hat{d}_L(n - M_L - i) \\ + G_H\hat{r}_H(n) + \sum_{i=1}^{N_{p_H}} \beta_{H,i}\hat{d}_H(n - M_H - i),\end{aligned} \tag{4.24}$$

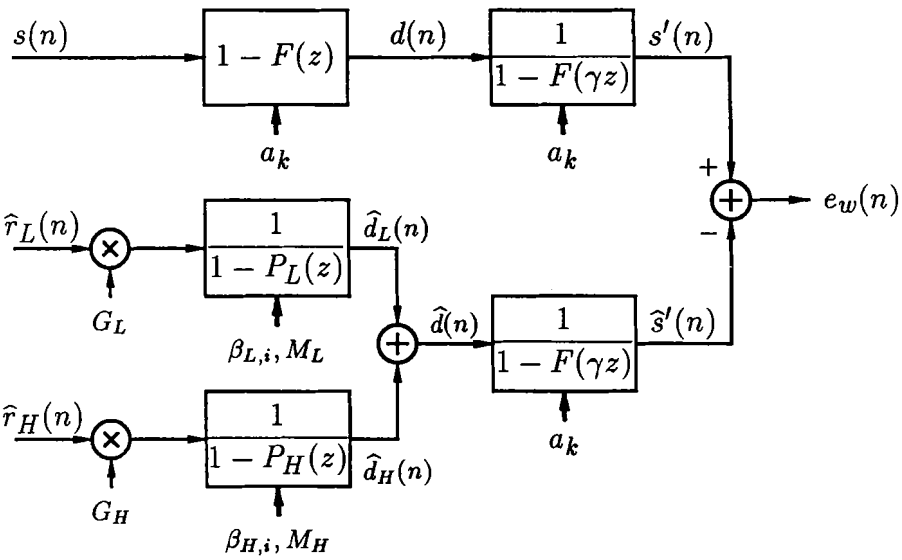
where N_{p_L} and N_{p_H} are respectively the number of pitch coefficients in the low and high band. The pitch lags M_L and M_H must both be larger than the pitch sub-frame size to prevent recursion. Then, $\hat{d}(n)$ can be viewed as a linear combination of all the known waveforms $\hat{r}_L(n)$, $\hat{r}_H(n)$, $\hat{d}_L(n - M_L - i)$ and $\hat{d}_L(n - M_H - i)$. Substituting Equation 4.24 into 4.5 yields:

$$\begin{aligned}\hat{s}'(n) = \sum_{k=-\infty}^{-1} \hat{d}_L(k)h'(n-k) + \sum_{k=-\infty}^{-1} \hat{d}_H(k)h'(n-k) \\ + \sum_{k=0}^{\infty} \hat{d}_L(k)h'(n-k) + \sum_{k=0}^{\infty} \hat{d}_H(k)h'(n-k).\end{aligned} \tag{4.25}$$

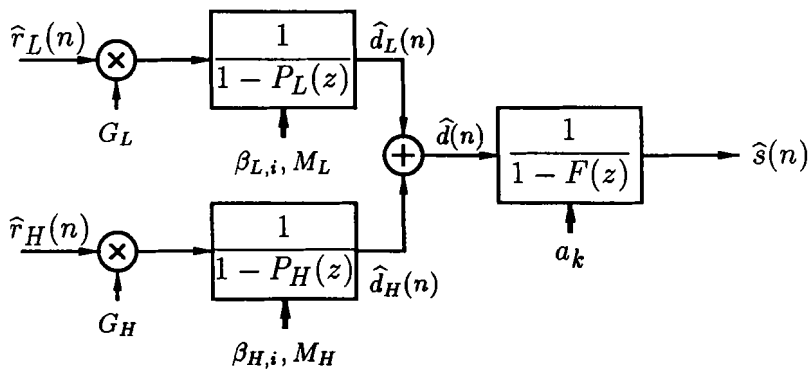
The anti-causal terms in the above equation are the zero-input responses of the bandwidth expanded formant synthesis filter and account for the initial conditions of each band at the pitch sub-frame boundaries. The impulse response $h'(n)$ is causal, and the upper limit in both causal terms summations can be set to $N - 1$. Defining



(a) residual band separation



(b) split–band optimization structure



(c) split–band synthesis structure

Fig. 4.5 Split–band RELP with pitch optimization.

the following terms,

$$x_L(n) = \sum_{k=0}^{N-1} \hat{r}_L(k)h'(n-k),$$
$$x_H(n) = \sum_{k=0}^{N-1} \hat{r}_H(k)h'(n-k), \tag{4.26}$$

and

$$y_{L,i}(n) = \sum_{k=0}^{N-1} \hat{d}_L(k-M_L-i)h'(n-k),$$
$$y_{H,i}(n) = \sum_{k=0}^{N-1} \hat{d}_H(k-M_H-i)h'(n-k), \tag{4.27}$$

the weighted error $e_w(n)$ can be expressed as:

$$e_w(n) = s^*(n) - G_L x_L(n) - \sum_{i=1}^{N_{p_L}} \beta_{L,i} y_{L,i}(n), \\ - G_H x_H(n) - \sum_{i=1}^{N_{p_H}} \beta_{H,i} y_{H,i}(n), \tag{4.28}$$

where $s^*(n)$ contains all the terms not subjected to optimization:

$$s^*(n) = s'(n) - \sum_{k=-\infty}^{-1} \hat{d}(k)h'(n-k). \tag{4.29}$$

Minimization is done in the mean-square sense. Differentiating Equation 4.12 with respect to the gains and coefficients of both the low and high bands yields a set of linear equations similar to Equations 4.13 and 4.14:

$$\frac{\delta\xi}{\delta G_L} = \sum_{n=0}^{N-1}\Big[2\Big[s^*(n) - G_L x_L(n) - G_H x_H(n) \\ - \sum_{i=1}^{N_{p_L}} \beta_{L,i} y_{L,i}(n) - \sum_{i=1}^{N_{p_H}} \beta_{H,i} y_{H,i}(n)\Big]\Big[-x_L(n)\Big]\Big] \\ = 0, \tag{4.30}$$

$$\frac{\delta\xi}{\delta G_H} = \sum_{n=0}^{N-1}\Big[2\Big[s^*(n) - G_L x_L(n) - G_H x_H(n) \\ - \sum_{i=1}^{N_{p_L}} \beta_{L,i} y_{L,i}(n) - \sum_{i=1}^{N_{p_H}} \beta_{H,i} y_{H,i}(n)\Big]\Big[-x_H(n)\Big]\Big] \\ = 0, \tag{4.31}$$

for $j = 1 \ldots N_{p_L}$,

$$\frac{\delta\xi}{\delta\beta_{L,j}} = \sum_{n=0}^{N-1}\Bigg[2\Big[s^*(n) - G_L x_L(n) - G_H x_H(n) - \sum_{i=1}^{N_{p_L}} \beta_{L,i} y_{L,i}(n) - \sum_{i=1}^{N_{p_H}} \beta_{H,i} y_{H,i}(n)\Big]\Big[-y_{L,j}(n)\Big]\Bigg] \tag{4.32}$$
$$= 0,$$

and for $j = 1 \ldots N_{p_H}$,

$$\frac{\delta\xi}{\delta\beta_{H,j}} = \sum_{n=0}^{N-1}\Bigg[2\Big[s^*(n) - G_L x_L(n) - G_H x_H(n) - \sum_{i=1}^{N_{p_L}} \beta_{L,i} y_{L,i}(n) - \sum_{i=1}^{N_{p_H}} \beta_{H,i} y_{H,i}(n)\Big]\Big[-y_{H,j}(n)\Big]\Bigg] \tag{4.33}$$
$$= 0.$$

When simplified, Equations 4.30 to 4.33 can be expressed in matrix form as $\Phi\mathbf{v} = \mathbf{b}$, where Φ, $\mathbf{v}$ and $\mathbf{b}$ are as follows†:

$$\Phi = \langle \mathbf{q}\mathbf{q}^{\mathrm{T}} \rangle, \tag{4.34}$$

where

$$\mathbf{q} = \begin{pmatrix} x_L(n) \\ x_H(n) \\ y_{L,1}(n) \\ \vdots \\ y_{L,N_{p_L}}(n) \\ y_{H,1}(n) \\ \vdots \\ y_{H,N_{p_H}}(n) \end{pmatrix}, \tag{4.35}$$

and

† For clarity, all summation symbols are left out. Thus $\langle x(n) \rangle$ refers to $\sum_{n=0}^{N-1} x(n)$.

$$\mathbf{v} = \begin{pmatrix} G_L \\ G_H \\ \beta_{L,1} \\ \vdots \\ \beta_{L,N_{p_L}} \\ \beta_{H,1} \\ \vdots \\ \beta_{H,N_{p_H}} \end{pmatrix}, \tag{4.36}$$

$$\mathbf{b} = \begin{pmatrix} \langle s^*(n)x_L(n)\rangle \\ \langle s^*(n)x_H(n)\rangle \\ \langle s^*(n)y_{L,1}(n)\rangle \\ \vdots \\ \langle s^*(n)y_{L,N_{p_L}}(n)\rangle \\ \langle s^*(n)y_{H,1}(n)\rangle \\ \vdots \\ \langle s^*(n)y_{H,N_{p_H}}(n)\rangle \end{pmatrix}. \tag{4.37}$$

This linear system is solved using the Cholesky algorithm. Since Φ may be ill-conditioned, the precautions described in Section 4.2.1 also hold for this split–band system.

Finally, the solution vector $\mathbf{v}$ depends on the pitch lags M_L and M_H since $y_{L,i}(n)$ and $y_{H,i}(n)$ are functions of M_L and M_H (see Eq. 4.27). The overall optimal solution is thus obtained through an exhaustive search of possible lag values within the allowed pre-defined lag range for each band.

4.3 Performance analysis

The optimization procedures developed in the previous section are tested without any parameter quantization. The test material used for the performance analysis consists of the first two sentences spoken by each speaker, as listed in Appendix A, for a total of 8 sentences. The goal of the tests is to verify the relative performances

of the full–band and split–band optimization procedures of Sections 4.2.1 and 4.2.2, and to assess the perceptual impact of the extra 4 kHz of bandwidth. For comparison purposes, the non-optimized RELP coder of Section 4.1 is also part of the test. The performance evaluation of the coders is based on Segmental SNR measures and on informal listening tests.

Mode	Optimization method	Preserved baseband (kHz)	Number of pitch taps
no1-1	none	1	1
no1-3	none	1	3
no2-1	none	2	1
no2-3	none	2	3
no4-1	none	4	1
no4-3	none	4	3
fb1-1	full	1	1
fb1-3	full	1	3
fb2-1	full	2	1
fb2-3	full	2	3
fb4-1	full	4	1
fb4-3	full	4	3
sb1-1	split	1	1
sb1-3	split	1	3
sb2-1	split	2	1
sb2-3	split	2	3
sb4-1	split	4	1
sb4-3	split	4	3

Table 4.1 Enhanced wideband RELP test configurations. For the split–band optimization configuration, the number of pitch taps is the same in both bands.

Throughout this section, the coder configurations are referred to by optimization method, preserved residual baseband and number of pitch parameters. The various

configurations are listed in Table 4.1. Also, all configurations use 16 poles for the LPC analysis, updated at 64 Hz (frame size of 250 samples), while the pitch parameters are updated at 320 Hz (sub-frame size of 50 samples). The lag is allowed to vary between 51 and 320 samples (corresponding to pitch periods ranging from 3.2 ms to 20 ms). In the case of split–band optimization, the lag is the same for both bands (i.e. $M_L = M_H$). This is a practical choice, as separate lag values require a large amount of calculations for each sub-frame. This sub-optimal approach has minimal effects of the performance. Simulations with separate lag optimization showed that most of the time, both lag values were either close to each other, or close to a common multiple. Finally, distinct optimal gain values are used for each band in the split–band configuration.

4.3.1 SegSNR performance

Figure 4.6 shows the SegSNR performance of the coders as a function of the preserved residual baseband. The solid curves are for coders with 1 pitch tap while the dotted curves are for coders with 3 pitch taps. The first observation is that optimization always improves the SegSNR of the regenerated speech. This is most evident when only 1 kHz of baseband is preserved: for the 1 pitch tap case, the SegSNR jumps from 6.3 dB for the non-optimized coder (None–1 curve) to 11.0 and 14.1 dB for the full and split–band coders respectively (Full–1 and Split–1 curves). These improvements are not as good in the 3 pitch tap case. This seems to indicate that within 1 kHz of preserved baseband residual, there may not always be enough pitch information for the optimization procedures to efficiently track and model the

pitch characteristics, and that there is little advantage in using 3 pitch taps instead of 1. However, as more baseband residual is preserved, the use of 3 pitch taps improves the SegSNR of the full and split–band optimization approaches by as much as 4 dB.

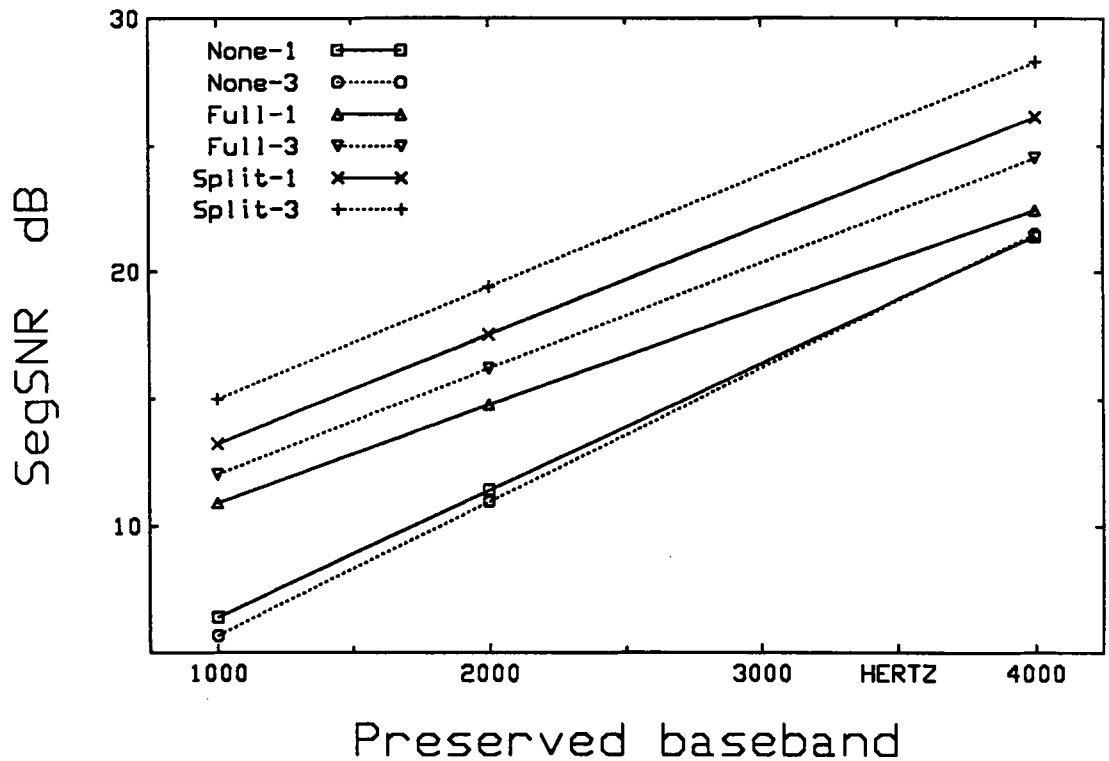


Fig. 4.6 SegSNR performance of the enhanced wideband RELP coders.

Second, for any number of pitch taps, the SegSNR of the split–band coders is always 3 to 4 dB higher than that of the full–band coders. This improvement confirms the original assumption that the optimized full–band parameters may no longer be optimal with respect to the preserved baseband (Sec 4.2.2). Indeed, there are times when the reconstructed speech baseband of the full–band optimization approach is distorted. An example of this is shown in Figure 4.7, where 2 kHz of baseband residual is preserved and 1 pitch tap is used. The full–band method (middle trace) introduces

distortion not found in the non-optimized method (top trace) (i.e. around 1.5 kHz). The split–band method (bottom trace) does not exhibit such baseband distortions. It also does a better job at reproducing the upper band than the other two approaches.

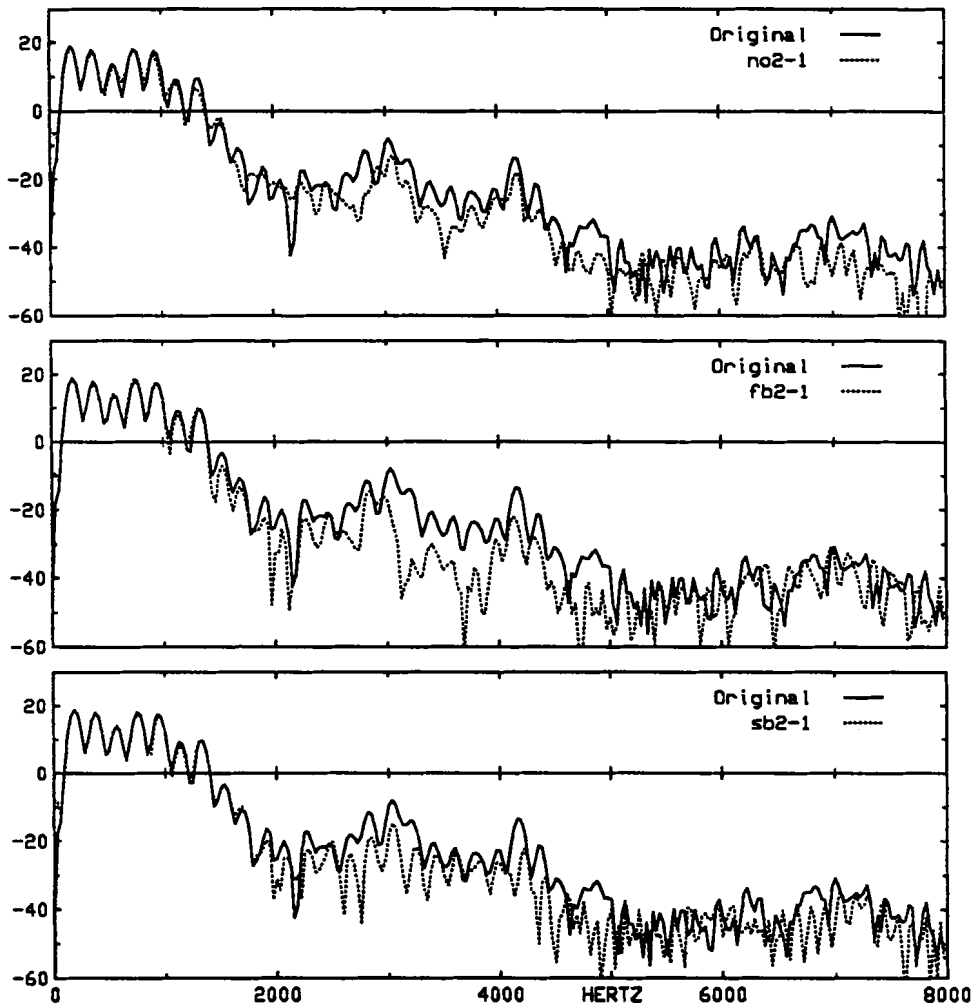


Fig. 4.7 Reconstructed speech spectrum for the enhanced wideband RELP coders. In all 3 cases, 2 kHz of baseband residual is preserved. The top trace shows the non-optimized approach with a single pitch tap. The middle trace shows the full-band optimization approach with a single pitch tap, whereas the bottom trace shows the split-band approach with a pitch tap in both the low and high band.

Although the preserved baseband residual is not explicitly coded, it is stored in

a file at some point in the simulation process. This storage involves a floating-point to integer conversion equivalent to 16-bit uniform quantization. This induces some baseband speech degradations found below 2 kHz in the non-optimized method (top trace).

4.3.2 Subjective performance

The listening tests are consistent with the SegSNR measures. First, for all methods, there is very little difference between the reproduced and original speech when 4 kHz of baseband residual is preserved. However, when 2 kHz is preserved, the split–band method clearly outperforms the other two pitch configurations. In particular, the lower frequencies are better rendered, and, although still a bit hollow, the speech is not metallic as in the full–band method. Here the use of 3 pitch taps instead of 1 makes a noticeable positive difference. Finally, when only 1 kHz of baseband is kept, there is little difference between the split–band and full–band method. Both yield hollow, slightly metallic sounding speech, regardless of the number of pitch taps. Still, both are definitely better than the non-optimized method.

Overall, these listening tests confirm the benefits of pitch re-optimization. They also prove that it is possible to code a wideband speech signal with less attention given to the upper band. This is a crucial observation. In essence, it shows that if a sufficient amount of low frequencies, say 4 kHz, are very well coded, then a sub-optimal excitation waveform can be used to regenerate the high band with little perceived distortion. This concept was originally proposed in the basic RELP model,

but with medium success due to the limited amount of preserved baseband (i.e. less than 2 kHz). In a wideband context however, its effectiveness becomes obvious.

4.4 Baseband Residual Coding

Preserving 4 kHz of baseband residual amounts to coding a signal sampled at 8 kHz. If the overall operating rate is to be 16 kbits/sec, the baseband transmission, including the gain factor, should not take more than 8000 bits/sec in order to leave enough bits to code the remaining parameters (i.e. LSF's, pitch coefficients and lag). This would require a baseband coding scheme operating at less than 1 bit/sample. No scalar method can perform at such a low rate with acceptable quality. One must then turn to vector quantization methods (VQ), where blocks of samples are coded rather than the individual samples [1]. Typically, a codebook of size 2^B, with codewords of length N is used. When coding a block of N samples, the codebook is searched and the closest codeword, usually in the mean-square sense, is selected and its index transmitted (i.e. B bits). The number of bits/sample is then B/N, a ratio ranging between 0.25 and 0.5 for acceptable quality.

Some form of VQ will be used to code the baseband residual. If a codebook is used, then all possible excitation waveforms are known *a-priori*. This is a clear advantage over scalar residual quantizing, as one can now try each possible codeword and retain the one that produces the best speech rather than the one that best matches the original residual. This can be done for both full and split–band optimization. In essence, this is CELP. The waveform selection process and possible codebook designs are discussed in the next chapter.

Chapter 5 Wideband CELP Coding

The basic CELP model presented in Section 2.3 can be adapted to operate in a wideband environment. As is, it simply follows the full–band optimization procedure of Section 4.2.1. However, it can be modified to operate as a split–band CELP coder, based on the split–band optimization procedure of Section 4.2.2. In this case, both the lower and upper band excitation residuals $\hat{r}_L(n)$ and $\hat{r}_H(n)$ are selected from codebooks.

In this chapter, the first section presents full–band implementations while split–band systems are studied in the second section. In both cases, parameter design and selection issues are treated first. Then, simulation results are presented for varying operating conditions. The goal here is to study the effects of some operating parameters, and to try to obtain the set of parameters that yield the best reconstructed speech. The coding without quantization should ideally be transparent. To this effect, no quantization other than that introduced by the codewords themselves is considered. Finally, the last section compares the best full–band and split–band methods subjected to operating rate constraints.

5.1 Full–band CELP

A number of alternative full–band CELP implementations are considered. In particular, the number of LPC coefficients a_k is kept constant at 16. The sections below describe the other operating parameters more precisely. Simulation results follow.

5.1.1 Frame and sub-frame sizes

The frame and sub-frame sizes control the update rate of all the coding parameters. Two approaches are taken. First, the frame and sub-frame sizes are respectively set to 250 and 50 samples (250:50 mode). For 16 kHz sampling, this corresponds to update rates of 64 Hz and 320 Hz respectively. In a second case, the frame and sub-frame sizes are respectively set to 320 and 40 samples (320:40 mode). For 16 kHz sampling, this corresponds to update rates of 50 Hz and 400 Hz respectively. In both cases, these update rates are typical for narrowband CELP systems.

5.1.2 Codeword design and selection

For this study, codebook sizes between 8 and 1024 codewords are used. The codeword length is always the same as the sub-frame size. When operating in full–band mode, a single full–band excitation drives the cascade of formant and pitch synthesis filters. The codebook consists of normalized *iid* Gaussian sequences. The optimal codeword is selected by solving the linear system of equations presented

in Section 4.2.1 (Eq. 4.18) for each codeword entry, and keeping the index of the codeword that yields the smallest error energy (Eq. 4.12).

5.1.3 Lag estimate

The maximum lag value is set at 20 ms, while the minimum value is always set at 1 sample larger than the sub-frame size. The lag selection process is either optimal or sub-optimal. The pseudo-code in Figure 5.1 shows the difference between the optimal (Fig. 5.1a) and sub-optimal (Fig. 5.1b) lag selection methods.

```
let ξ_min = ∞
for M = Minlag, Maxlag
    for each codeword i
        Solve the system Φv = b
        Compute the error energy ξ
        if ξ < ξ_min
           the optimal lag is M
           the optimal codebook index is i
           the optimal solution vector is v
           let ξ_min = ξ
        endif
    endfor
endfor
```

(a) optimal method

Fig. 5.1 Full–band lag and codeword selection.

In the optimal case, for each possible lag value (between $Minlag$ and $Maxlag$, the linear system of equations must be solved for each codebook entry (i.e. 2^B possible codewords). For large codebooks, this exhaustive nested search is too computationally intensive.

```
let ξ_min = ∞
let G = 0
for M = Minlag, Maxlag
    Solve the system Φv = b
    Compute the error energy ξ
    if ξ < ξ_min
       the optimal lag is M
       let ξ_min = ξ
    endif
endfor

let ξ_min = ∞
for each codeword i
    Solve the system Φv = b
    Compute the error energy ξ
    if ξ < ξ_min
       the optimal codebook index is i
       the optimal solution vector is v
       let ξ_min = ξ
    endif
endfor
```

(b) sub-optimal method

Fig. 5.1 Full–band lag and codeword selection.

In the sub-optimal approach, the lag selection is decoupled from the codeword selection. In other words, the lag is found before selecting the optimal codeword. The sub-optimal lag search is is done by forcing the gain factor G to zero, thus eliminating any contributions from the current excitation. The resulting reduced system of equations is then solved for all possible lag values within the pre-defined lag range. In essence, this amounts to letting the pitch synthesis filter *free-wheel* (or self-excite) with the past regenerated formant residual (see Fig. 4.5). This approach is only slightly sub-optimal, and eliminates the computational burden induced by nesting exhaustive lag and codebook searches. The loss in performance is small [18] since

the contributions to the pitch structure primarily come from the past regenerated formant residual and not from the current excitation.

5.1.4 Pitch prediction order

The number of pitch coefficients is either 1 or 3. When quantization is on, the computed optimal pitch coefficients are quantized before the error energy (Eq. 4.12) is calculated. The quantization noise is therefore accounted for within the optimization. This is optimal only if there is a single coefficient per band (regardless of the mode). A fully optimal, but impractical solution would involve nested searches through all possible quantized values of each pitch coefficient. Here, it is assumed that the coefficients are quantized independently.

5.1.5 Gain

When quantization is enabled, a differential quantizer with a leaky predictor (i.e. DPCM with 1 tap prediction, tap=0.9) is used to code the difference between successive sub-frame gain magnitudes [18]. An extra bit is required for the sign of the gain magnitudes. This can be seen as doubling the codebook size. Also, as for the pitch parameters, the quantized gain value is used to compute the error energy. This ensures the overall best solution under the constraints of parameter quantization by including any distortion induced by the gain quantizer.

5.1.6 Performance analysis

The performance of various full–band implementations is studied. The material

used for the analysis consists of the first sentence spoken by each speaker, as listed in Appendix A, for a total of 4 sentences.

For given frame and sub-frame sizes, the performance of the coder improves with the size of the codebook and the number of pitch taps. The simulation results are listed in Table 5.1.

Codebook	250:50 mode		320:40 mode	
size	1 tap	3 taps	1 tap	3 taps
8	10.28	11.52	11.24	12.58
16	10.76	12.03	11.28	13.34
32	11.54	12.77	12.54	14.22
64	11.94	13.25	13.11	14.82
128	12.49	13.85	13.47	15.19
256	12.82	14.11	14.16	15.60
512	13.31	14.59	14.62	16.23
1024	13.49	14.63	14.95	16.68

Table 5.1 Full–band SegSNR performance (dB).

In both update modes, there is a relatively constant 1.25 dB to 1.7 dB improvement in SegSNR when using 3 pitch coefficients instead of 1, regardless of the codebook size. Perceptually, this improvement is mostly felt for small codebooks, but becomes less noticeable for codebooks of 512 and 1024 waveforms. Another interesting observation is that the SegSNR only increases by 0.3 dB to 0.5 dB with each doubling of the codebook size. These increases saturate for large codebooks (i.e. from 512 to 1024), perhaps indicating that beyond a certain codebook size, other parameters, such as the number of pitch taps or update rates, have more influence.

In general, listening tests indicate that the 320:40 update mode is better than the 250:50. For the 250:50 mode, most simulations yield reconstructed speech that contains some noticeable distortions, even with large codebooks and 3 pitch taps. At times, the coded speech sounds a little hoarse and some high frequency noise is perceivable. The 320:40 mode can, however, achieve near-transparent coding, although some high frequency distortions are still noticeable for certain sentences, especially near fricatives.

In summary, the above results show that faster update rates always yield better results for any codebook size. For smaller codebooks, 3 pitch parameters are better than 1. Finally, for large size codebooks, doubling the codebook size only marginally increases the quality of the coded speech.

5.2 Split–band CELP

When the coder operates in split–band mode, extra parameters need to be coded. Most parameters are selected as for the full–band coder. In particular, the frame and sub-frame update sizes are either set to 250:50 samples, or to 320:40 samples. Also, the number of poles is kept fixed at 16. The following sections describe other implementation details specific to split–band coding.

5.2.1 Lag estimate

As for the full–band case, the lag selection process is decoupled from the codeword selection. Also, the same lag value is used for both bands (i.e. $M_L = M_H$). This

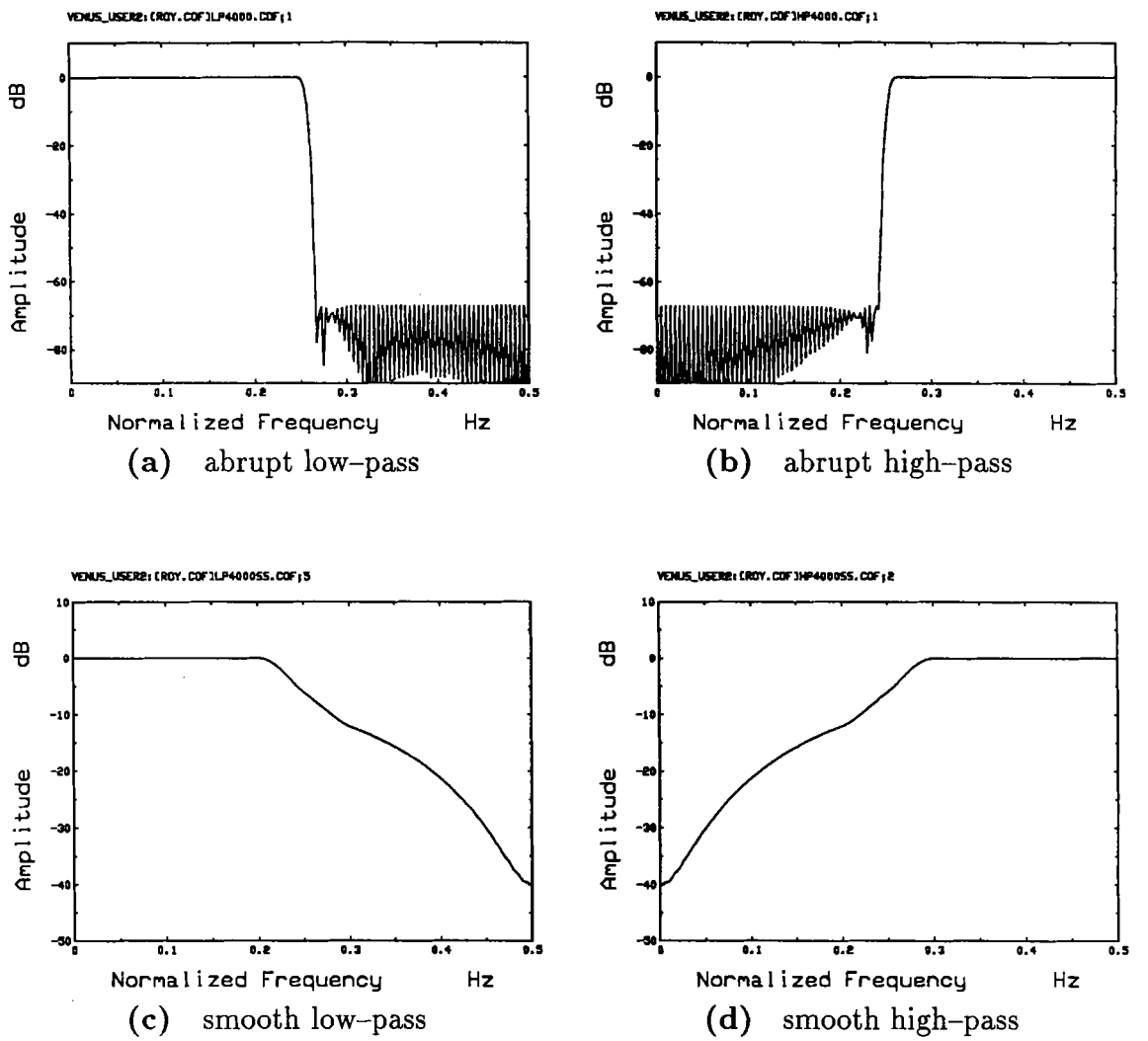
choice is a practical one, as lag coding usually requires 7 to 8 bits per sub-frame. Results in the previous chapter showed that a common lag value for both bands does not reduce the perceived quality of the reconstructed speech. Thus, the extra bits required to transmit a separate high–band lag value are better spent on other parameters. Finally, the maximum lag value is set at 20 ms, while the minimum value is always set at 1 sample larger than the sub-frame size.

5.2.2 Codeword design and selection

When operating in split–band mode, separate excitations are required for each band and thus, a low and a high–band codebook are used. For this study, codebook sizes between 8 and 1024 codewords are used. The codeword length is always the same as the sub-frame size.

The codebooks can either be normalized *iid* Gaussian sequences (as in the full–band case), or band-limited normalized Gaussian sequences. Band-limiting the codebooks is done at design time by filtering a Gaussian sequence with a low or high–pass filter. The idea of band-limiting the codebooks is based on the results obtained in the previous chapter (where $\hat{r}_L(n)$ and $\hat{r}_H(n)$ were respectively restricted to the low and high band), and is an attempt to prevent each band from adversely affecting the other. Thus, codebooks can either be of a full, low or high–pass nature.

Two types of band-limiting are considered and are shown in Figure 5.2. The first one is abrupt (Fig. 5.2(a)): both the low and high–pass filters are complementary high-order (201 taps) FIR filters with a stop–band attenuation of 70 dB. The second type is smoother (Fig. 5.2(b)): both the low and high–pass filters are complementary



(a) abrupt low–pass (b) abrupt high–pass

(c) smooth low–pass (d) smooth high–pass

Fig. 5.2 Codebooks band-limiting filters.

high-order (201 taps) FIR filters with a gradual transition band. In both cases, 4 kHz is chosen as the cutoff frequency. This choice is based on the previous chapter, and serves as a good boundary between the respective resolution (i.e. codebook sizes) given to each band.

Given low and high codebooks of 2^{B_L} and 2^{B_H} codewords respectively, the optimal split–band codewords can be found in an optimal or sub-optimal way. The pseudo-code in Figure 5.3 shows the difference between the optimal (Fig. 5.3(a) and

sub-optimal (Fig. 5.3(b) codewords selection methods. In both methods, the common lag value is determined sub-optimally, as in the full–band structure.

```
let ξ_min = ∞
for each low band codeword i_L
    for each high band codeword i_H
        Solve the system Φv = b
        Compute the error energy ξ
        if ξ < ξ_min
           the optimal low index is i_L
           the optimal high index is i_H
           the optimal solution vector is v
           let ξ_min = ξ
        endif
    endfor
endfor
```

(a) optimal method

Fig. 5.3 Split–band codewords selection.

In the optimal method (i.e. "SB–OPT"), both codebooks are exhaustively searched : for each low–band codeword, the high–band codebook is searched and the indices i_L and i_H that minimize the error energy (Eq. 4.12) are selected. This is optimal but computationally intensive due to the nested searches.

In the sub-optimal method (i.e. "SB–SUB"), both codebooks have the same size (i.e. $B_L = B_H$). The search is conducted with respect to one of the codebooks only. A single index is transmitted and shared by the two codebooks. Since most of the error energy ξ comes from the low–band contribution and since the optimization is done by minimizing ξ, one can view this as an exhaustive search method for the low codebook. Sharing the optimal index then amounts to selecting an arbitrary high–band excitation. This is somewhat analogous to what was done in the previous chapter

```
let ξ_min = ∞
for each low band codeword i_L
    Let i_H = i_L
    Solve the system Φv = b
    Compute the error energy ξ
    if ξ < ξ_min
       the optimal low index is i_L
       the optimal high index is i_H
       the optimal solution vector is v
       let ξ_min = ξ
    endif
endfor
```

(b) sub-optimal method

Fig. 5.3 Split–band codewords selection.

(i.e. simple spectral folding essentially yielded an arbitrary high–band excitation). In SB–SUB, the upper band excitation comes in at no extra bit cost (i.e. zero bit codebook).

5.2.3 Pitch prediction order

In split–band mode, there is either one pitch coefficient in each band (1:1 mode) or three in the low and 1 in the high (3:1 mode). The optimality conditions for quantization are the same as in the full–band case.

5.2.4 Gain

In split–band mode, either a single gain value common to both bands is computed, or distinct values G_L and G_H are computed for each band. The optimality conditions for quantization are the same as in the full–band case.

5.2.5 Performance analysis

The effects of band-limiting the codebooks are first studied, followed by gain coding effects. The SB–OPT and SB–SUB methods are then compared. Finally, the number of pitch taps and frame update rates is varied.

5.2.5.1 Codebook band-limiting effects

Band-limiting the codebooks affects the coder performance. The codebooks are of either full, low–pass or high–pass nature (designated as F, L and H). The transition bands are either abrupt or smooth. For simplicity, the other operating parameters were held constant. The test configurations and their respective SegSNR are listed in Table 5.2.

Mode	Transition	Low-book	High-book	SegSNR
F–F	*	Full	Full	12.52
L–H	Abrupt	Low	High	12.26
F–H	Abrupt	Full	High	12.02
L–F	Abrupt	Low	Full	12.52
L–H	Smooth	Low	High	12.26
F–H	Smooth	Full	High	12.01
L–F	Smooth	Low	Full	12.59

Table 5.2 Band-limiting configurations and SegSNR (dB). These results are for the SB–SUB method, with a 250:50 update mode, a 1:1 pitch prediction mode, a common lag and gain value for both bands and codebook sizes equal to 32.

First, the SegSNR performances are nearly all identical. Perceptually however, the F–F and F–H methods sound better. Both L–H configurations suffer mainly from

metallic noises and low-level distortions, while both L–F configurations, though they yield the highest SegSNR's, sound slightly muted in the high frequencies.

The F–F and F–H methods are further examined in Figure 5.4. The figure shows comparative frequency plots for the same frame of female speech. Overall, the abrupt F–H method offers the best harmonic match. The F–F method displays some harmonic discontinuities between 2 and 3 kHz that are not as noticeable in the abrupt F–H and smooth F–H method. Below 2 kHz however, the smooth F–H method shows distorsions which are probably induced by the leading edge of the frequency response of the high–band excitation. It thus seems advantageous to completely prevent the high–band excitation from contributing to the low–band regenerated speech. To this effect, only the abrupt F–H method is retained for further testing.

5.2.5.2 Gain effects

In the previous section, all the simulations were done with a single optimal gain value common to both bands. In this section, separate low and high–band optimal gain values are used, and the codebook sizes are varied from 8 to 1024 vectors . All other parameters remain the same.

The use of a separate gain parameter for the high–band improves the SegSNR figures by 0.2 to 0.75 dB depending on the codebook sizes. Perceptually, this improvement is usually difficult to notice, although it seems to reduce the level of hiss in some cases. The plots in Figure 5.5 show frequency responses for a frame of male and a frame of female speech. The codebooks in this case contain 32 codewords.

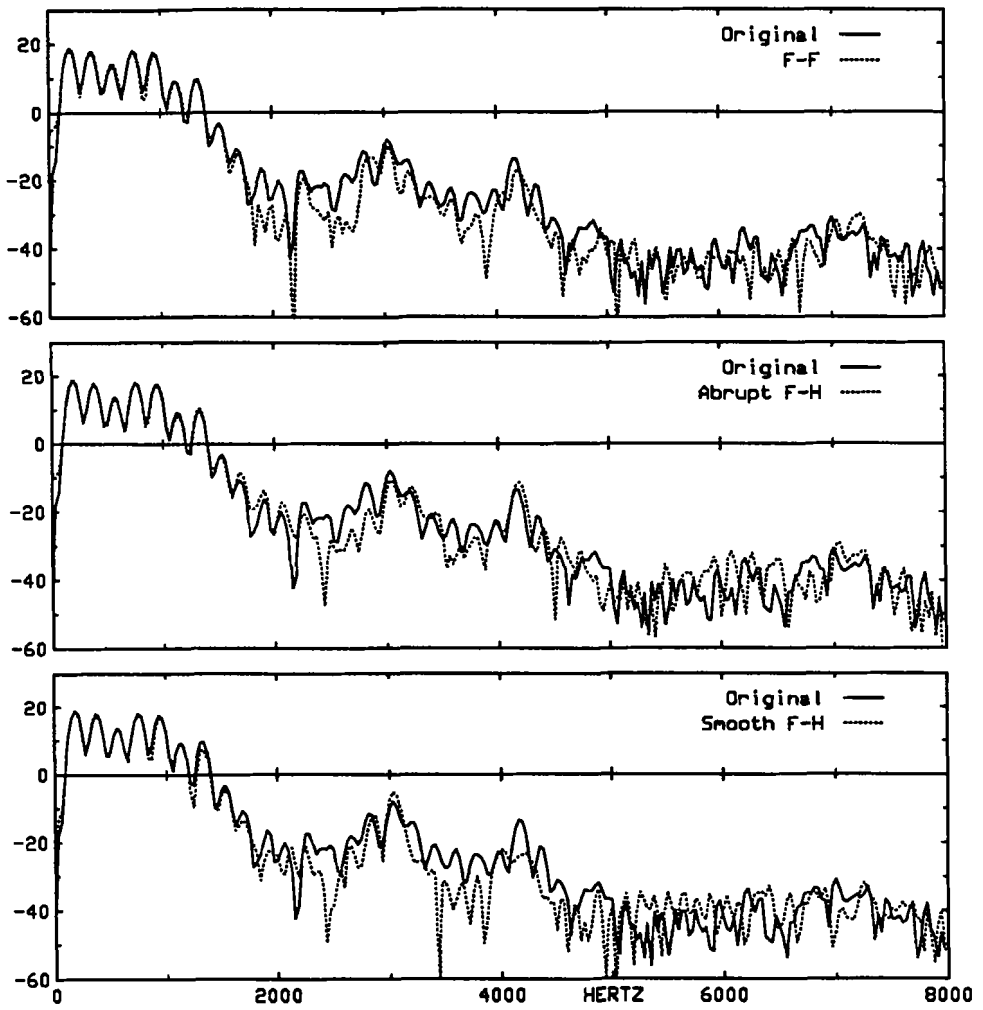


Fig. 5.4 Comparative speech spectrums for band-limited codebooks.

The first observation is that separate gain control yields a better spectral match throughout the band. In particular, the harmonics in the 1 to 4 kHz range are better modeled. In the common gain mode, the computed optimal gain compromises between the needs of both the low and high-band excitations. There is no such compromise in the separate gain control mode. When using F-H Abrupt codebooks, the low gain G_L affects the whole band, whereas the high gain G_H only affects the high band. It therefore seems advantageous to use separate gain control, but only

Codebook size	Gain	
	common	separate
8	10.82	11.04
16	11.42	11.75
32	12.00	12.30
64	12.55	12.86
128	13.04	13.50
256	13.31	13.80
512	13.63	14.02
1024	13.86	14.63

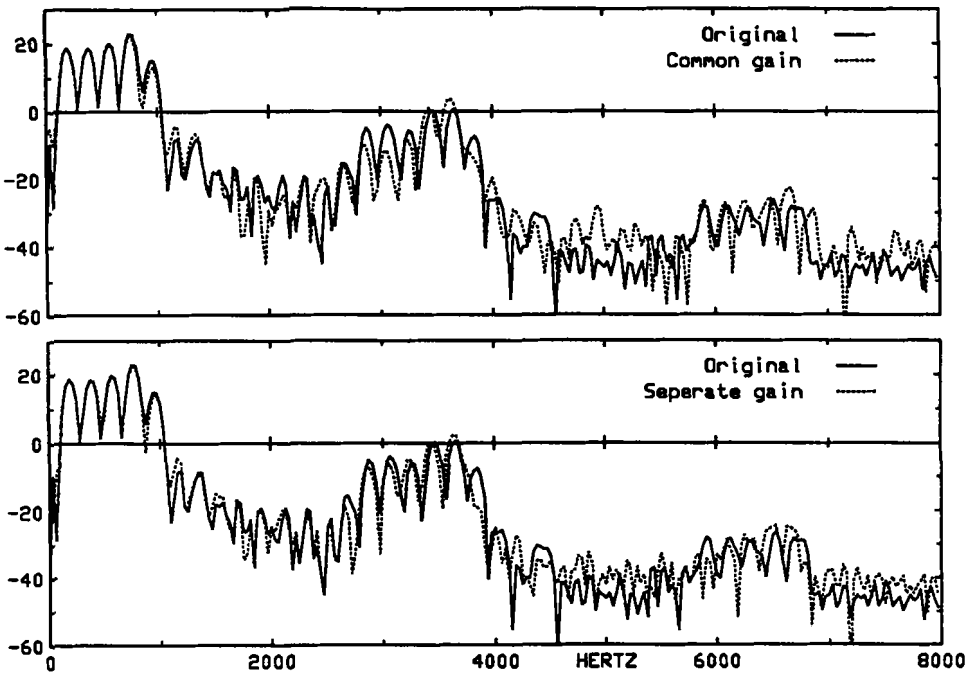
Table 5.3 Optimal gain selection and SegSNR (dB). These results are for the SB–SUB method, with a 250:50 update mode, a 1:1 pitch prediction mode, a common lag, separate gain values for both bands and abrupt F–H codebooks.

if the chosen operating bit rate can support the extra parameter G_H. Otherwise, a common gain scheme should prove adequate. For comparison purposes, the separate gain control approach is retained.

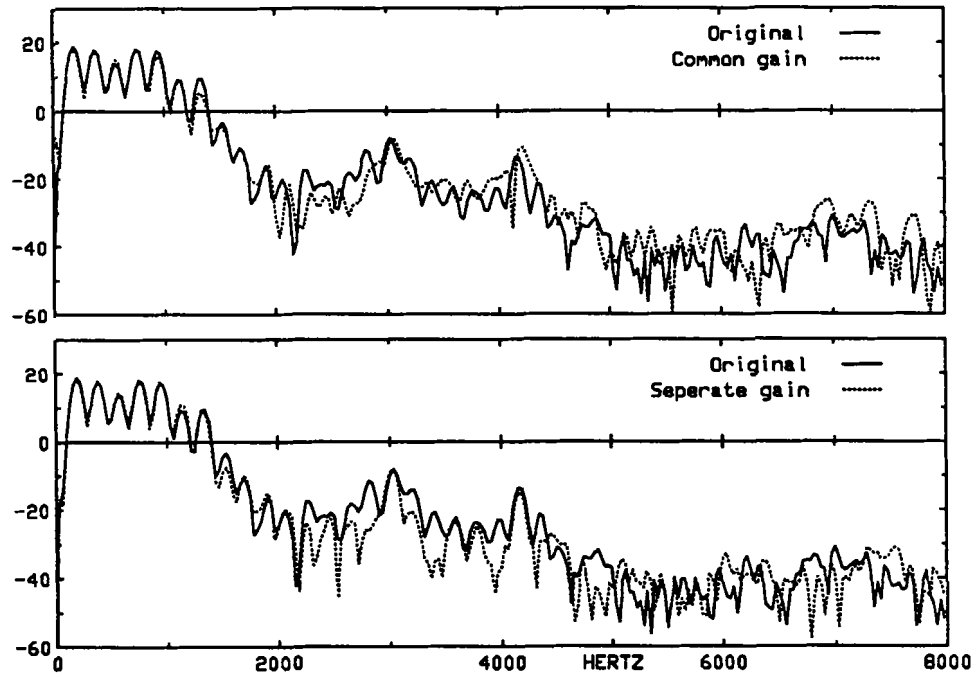
5.2.5.3 Codewords selection effects

In this section, the SB–OPT and SB–SUB codewords selection methods are compared. Since the SB–OPT method requires bits for both the low and high–band index, the comparison is restricted to a maximum of 10 bits for excitation modeling for either method. The choice of 10 bits is also a practical one, as the SB–OPT simulations tend to be computationally heavy due to the nested low and high–band exhaustive searches. The SegSNR results for both methods are listed in Table 5.4.

The simulation results indicate that the SB–SUB method only yields 0.26 dB less SegSNR than the best possible SB-OPT combination for 10 bits. The plots in



(a) male speech frame



(b) female speech frame

Fig. 5.5 Comparative speech spectrums for common and separate optimal gain control.

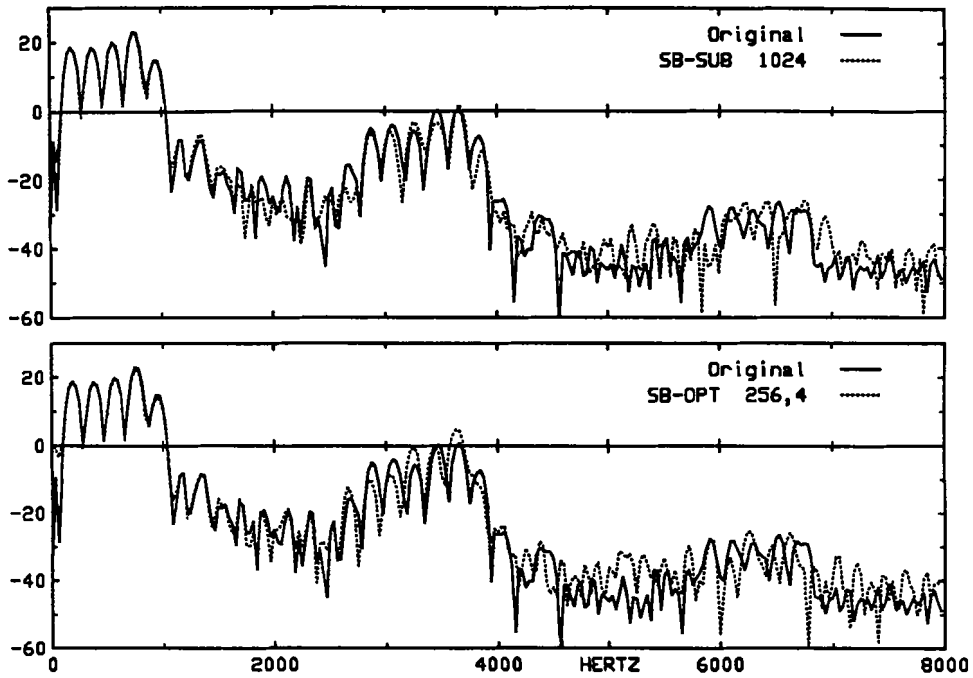
Mode	Codebooks sizes Low	High	SegSNR
SB–SUB	1024	–	13.86
SB–OPT	32	32	12.93
	64	16	13.47
	128	8	13.42
	256	4	13.86
	512	2	14.06
	1024	1	14.12

Table 5.4 Optimal and sub-optimal codewords selection and SegSNR (dB). These results are for a 250:50 update mode, a 1:1 pitch prediction mode, a common lag, separate gain values for both bands and abrupt F–H codebooks.

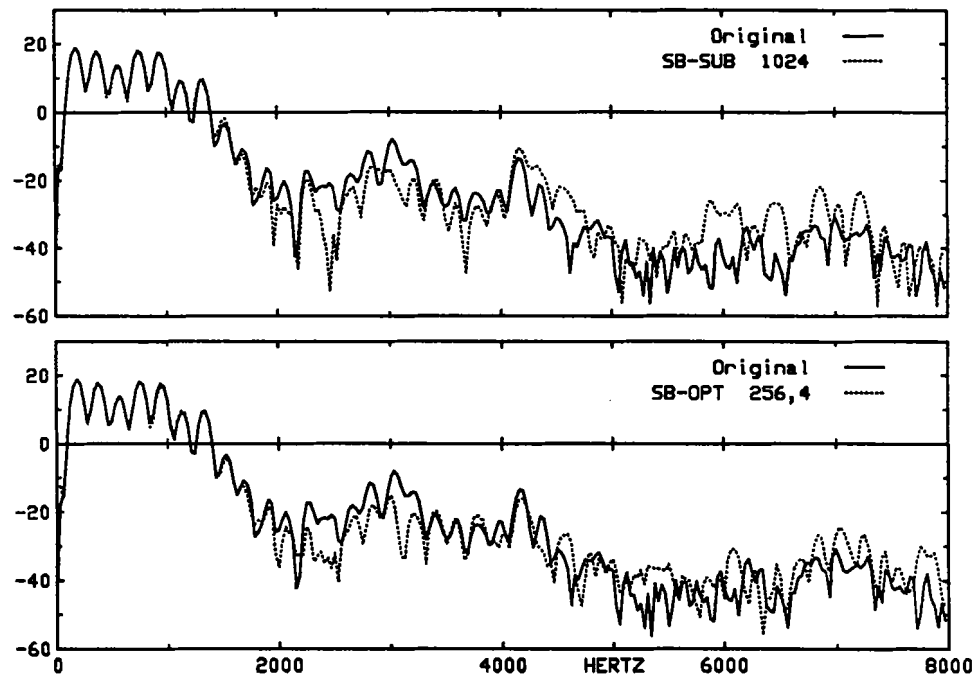
Figure 5.6 show frequency responses for a frame of male and a frame of female speech. In this case, the SB–OPT method has 256 low codewords and 4 high codewords. The SB–OPT seems to yield a slightly better harmonic structure than the SB-SUB method, especially for the female speech frame. However, listening tests show that the little differences between the two methods do not favor one over the other. Because of its simplicity and effectiveness, the sub-optimal high-band codeword selection method is therefore retained for further testing.

5.2.5.4 Pitch prediction order and update rate effects

In the previous sections, all simulations were done with the 1:1 pitch prediction mode and the 250:50 parameter update mode. This section presents results for the both pitch prediction modes and both parameter update modes. The SB–SUB method is used with abrupt F–H codebooks and separate gain values. The codebook



(a) male speech frame



(b) female speech frame

Fig. 5.6 Comparative speech spectrums for optimal and sub-optimal high band codeword selection.

Codebook	250:50 mode		320:40 mode	
size	1 tap	3 taps	1 tap	3 taps
8	10.82	12.24	12.61	14.11
16	11.42	13.05	13.21	14.72
32	12.02	13.72	13.99	15.31
64	12.55	14.12	14.43	15.81
128	13.04	14.48	14.94	15.93
256	13.31	14.92	15.14	16.76
512	13.63	15.35	15.75	17.04
1024	13.86	15.73	15.93	17.31

Table 5.5 Split–band SegSNR performance (dB). These results are for the SB–SUB method with separate gain values for both bands and abrupt F–H codebooks.

sizes are varied from 8 to 1024 codewords.

The results in Table 5.5 show the benefits of using more pitch taps and faster update rates. The 3:1 pitch prediction mode is clearly better than the 1:1 mode, regardless of the update mode used. Perceptually, this is only really noticeable for small codebook sizes. Listening tests also indicate that all 250:50 configurations still have some kind of noticeable high frequency distortions. However, these distortions tend to disappear in the 320:40 update mode. Indeed, near-transparent coding is achieved for large codebooks in either the 1:1 or 3:1 pitch prediction mode.

Another interesting observation is that the 1:1 pitch prediction mode in the 320:40 update mode is better than 3:1 in 250:50. Here the faster sub-frame rate more than compensates for the lower number of pitch taps. From an operating rate viewpoint, this is also advantageous. The sub-frame update rate is increased by 20% (i.e. from every 50 samples to every 40 samples). This rate increase for 2 pitch coefficients (i.e.

the 1:1 mode has 2 coefficients to transmit) still requires less bits/seconds than the 4 coefficients of the 3:1 mode.

This last observation is interesting in light of some of the work done on fractional pitch delay in narrowband systems. As opposed to integer pitch delays, fractional pitch delay estimates offer, for single tap pitch filters, a lag interpolating effect similar to that found in multiple tap pitch filters. Thus, better pitch filters can be achieved without increasing the overall bit rate. In this wideband case, the interpolating effects of the multiple tap pitch filter approach (i.e. the 3:1 mode) are more than compensated by a faster sub-frame update rate. Although this has not been done in this research, it would be interesting to determine if a fractional pitch delay method could further improve the performance of this wideband coder. It is suspected that the relative improvement may not be as important as that suggested in narrowband systems since the 16 kHz sampling frequency found in wideband systems already provides a "finer" integer lag estimate.

5.3 Comparison of Full and Split–band Wideband CELP

The previous two sections dealt with various implementation aspects of full and split–band wideband CELP coders. Based on the simulation results, the best full and split–band approaches were retained and are now compared while subjected to a maximum operating rate of 16 kbits/sec. This is still done however, with no quantization other than that introduced by the codeword selection. The operating rate calculations use estimated bit requirements for each parameter based on existing narrowband CELP implementations. The estimate for the LPC coefficients is based on

the results of Chapter 3. Two coder implementations are considered and listed in Table 5.6 and 5.7, both operating in the 320:40 update mode.

		320:40 mode	
Parameter	Bits	Update rate (Hz)	Bits/sec
LPC coefficients	48	50	2400
β_1	5	400	2000
β_2	3	400	1200
β_3	3	400	1200
gain G	6	400	2400
lag M	7	400	2800
codebook index	10	400	4000
		Total	16000

Table 5.6 Full-band coder configuration.

		320:40 mode	
Parameter	Bits	Update rate (Hz)	Bits/sec
LPC coefficients	48	50	2400
β_L	5	400	2000
β_H	3	400	1200
gain G_L	6	400	2400
gain G_H	4	400	1600
lag M	7	400	2800
codebook index	9	400	3600
		Total	16000

Table 5.7 Split-band coder configuration.

Both coders yield high quality reconstructed speech (unquantized parameters). In terms of SegSNR, the full-band implementation is about 0.5 dB higher than the

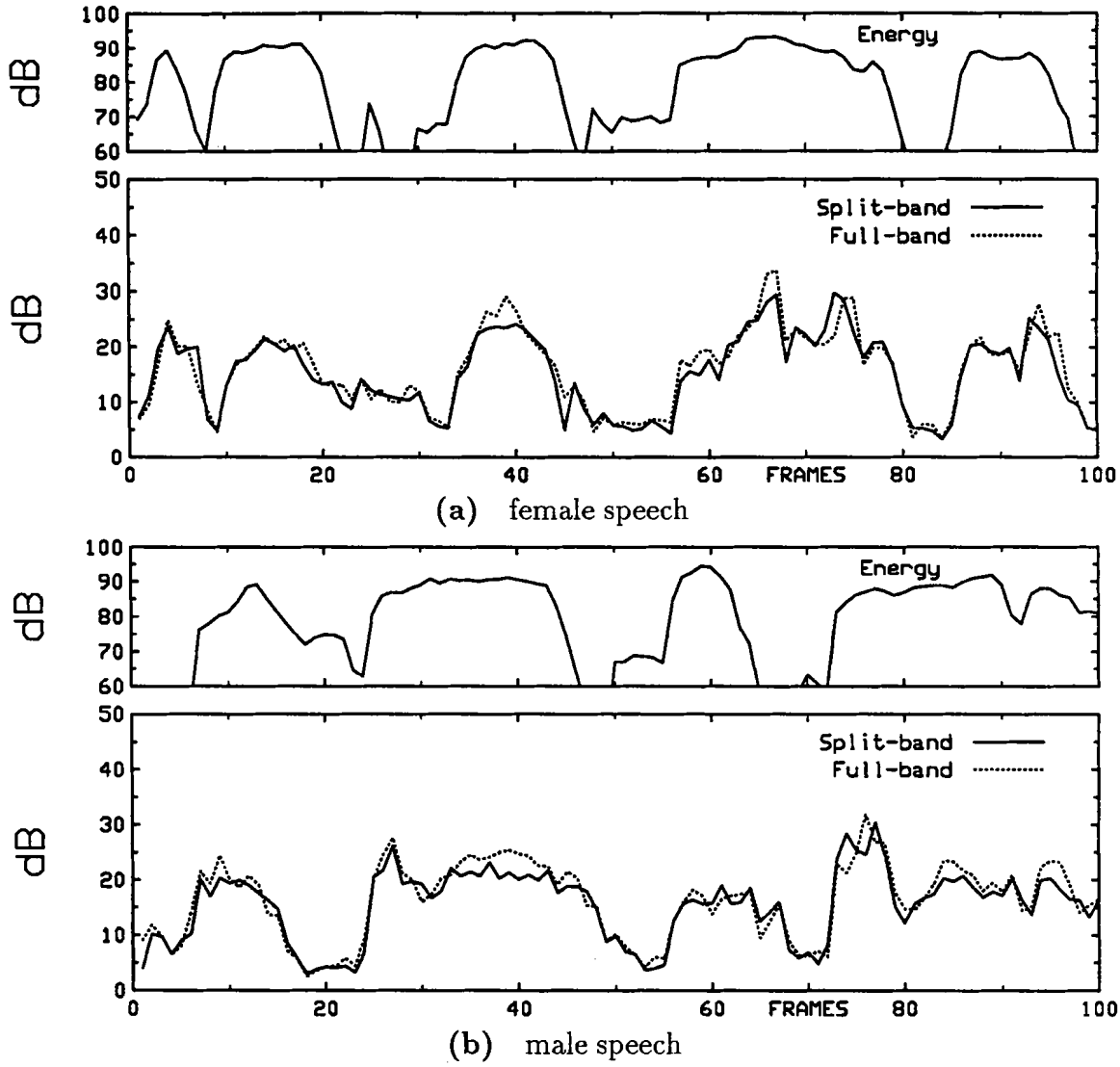


Fig. 5.7 Full versus split-band SegSNR.

split–band approach. The SegSNR tracks of Figure 5.7 show little overall difference between the two methods.

Perceptually, there are some cases where the full–band implementation suffers from a slight hollowness and from a certain hiss around fricatives. The split–band implementation does not exhibit these problems and generally produces a richer sound than the full–band method. In particular, it does a better job of reproducing the

baseband. This, in turn, seems to be an essential condition for the overall good reproduction of a wideband signal. Indeed, the results of the previous chapter showed no perceptual differences between the original speech and speech reconstructed from its original baseband and arbitrary high–band excitations. The split–band CELP structure proposed here follows this condition and emphasizes the low–band. The cost of the high–band is computed as follows: 1600 bits/sec for the extra gain factor G_H, 1200 bits/sec for the extra pitch tap β_H, 720 bits/sec for the LPC coefficients (assuming that 30% of the LPC bits are modeling the high–band), and finally 1400 bits/sec for the shared lag value. This adds up to 4920 bits/sec, or roughly 30% of the overall operating rate.

5.3.1 Comparison with a 16 kbits/sec narrowband coder

Finally, both coders were compared to a low-delay CELP narrowband coder operating at 16 kbits/sec [19]. For this comparison, the original wideband speech files were low–pass filtered at 3300 Hz, downsampled at 8 kHz and then processed by the narrowband coder. Informal tests were conducted with many different listeners to determine which of the two types of coders (narrowband *vs* wideband) was preferred. The wideband coders were always preferred over the narrowband one.

Although strictly subjective, this test clearly demonstrates that the reproduced wideband speech is of better quality. The extra bandwidth yields a "fuller" sound, and also greatly enhances the perception of fricative sounds. Even though the wideband CELP coders were not operating under full parameter quantization, these results nev-

ertheless indicate that for a potential operating bit rate of 16 kbits/sec, the wideband CELP coders can yield a clearer, richer sound than their narrowband counterparts.

Chapter 6 Conclusion

The objective of this thesis was to study the possible implementation of low–rate, analysis–by–synthesis wideband speech coders. For this, existing narrowband coding techniques were adapted to a wideband environment. In particular, known spectral envelope coding methods were extended to higher order systems, and special procedures for coding the higher frequencies found in a wideband signal were developed. In essence, simulation results showed that for an operating rate of 16 kbits/sec, the coded wideband speech sounds richer and clearer than its coded narrowband counterpart.

In the first part of the thesis, the wideband spectral envelope was coded using LSF's. The order of the formant prediction filter was set to 16, 18 and 20 poles. The LSF's were quantized using scalar techniques. The first method directly quantized each LSF independently (NUQ). This was inefficient and sometimes produced badly ordered LSF's since, as in narrowband systems, low order LSF's have wide and overlapping dynamic ranges. The second quantization method exploited this by coding the distance between adjacent LSF's rather than the LSF's themselves (DNUQ). Finally, the third method (TDNUQ) exploited the strong frame to frame correlation of LSF's by first quantizing the difference between odd LSF's from one frame to the

next, and then coding the even LSF's with respect to the coded odd LSF's. Both differential methods offer flexible control over the respective resolution given to each LSF, and thus, more emphasis was given to the low–band LSF's. Simulation results showed that the DNUQ approach performs better, and that at 50 bits/frame, no more than 16 poles need be used.

In the second part of this research, the basic RELP coder structure was enhanced by the insertion of a pitch prediction stage. This removed the harmonic structure from the formant residual, yielding a Gaussian–like, flat pitch residual spectrum. This signal was low–pass filtered and decimated. At the receiver, this signal was spectrally folded into the high–band (i.e. upsampled), yielding a sub-optimal, but spectrally flat, full–band excitation signal. Pitch optimization procedures were then developed to account for this sub-optimal excitation signal. In the full-band approach, an optimized set of pitch parameters was found for the whole band. In the split-band approach, separate low and high–band pitch parameters were found.

The simulations of the second part yielded two important results. First, when the low 4 kHz of pitch residual was left uncoded, an arbitrary excitation waveform could be used in the upper band with virtually no perceived degradation in the coded speech. In essence, well preserved low frequencies masked out the distortions induced by the sub-optimal high–band coding. Second, split-band optimization procedures performed better than full-band procedures by preventing each band from adversely affecting the other. The computed low and high band optimal pitch parameters accounted for distortions within their respective bands.

Finally, in the last section of this research, the baseband residual is coded using

stochastically populated codebooks. This amounted to wideband implementations of CELP, and the full and split–band procedures developed earlier were used. Simulations were performed with no other parameter quantization than that introduced by coding the baseband residual. Therefore, all operating rate estimates were based on the required number of bits for equivalent parameters in existing narrowband systems. In the full–band mode, the performance increased with the number of pitch taps and the parameter update rate, but tapered off with large codebooks (i.e. greater than 512 codewords). These observations also held for the split–band mode. Moreover, the best subjective performance was obtained when the low–band codebook contained full–band waveforms and the high–band codebook contained waveforms high–pass filtered at 4 kHz. As was demonstrated in the second part, the optimal high–band excitation could be chosen sub-optimally with little perceived distortion. Therefore, an exhaustive low–band optimal codeword selection mechanism was used for the low–band, while the high–band excitation was arbitrary (i.e. zero-bit codebook). Perceptually, there was little difference between this sub-optimal method and a fully exhaustive low and high–band codebook search.

As an overall conclusion, it seems possible to code a wideband speech signal at a rate of 16 kbits/sec. The results herein showed that although they greatly improve the perceived quality of a coded speech signal, the high frequencies found in a wideband signal need not be coded precisely.

The results obtained in this research remain preliminary. Because of time constraints, some coding aspects were left aside. The following section describes some of the issues that need to be investigated in greater depth.

6.1 Recommendations for Future Research

The first recommendation deals with LSF coding. LSF's offer a simple and flexible spectral envelope representation. The scalar LSF quantization techniques presented in this research may not be the most efficient. In narrowband systems, vector quantization (VQ) techniques can code 10 pole models with less than 30 bits per frame. It therefore seems reasonable to foresee wideband VQ methods operating at 40 bits per frame or less for systems with 16 or more pole. In particular, split-VQ techniques offer flexible resolution over the band by dividing the LSF's into 2 or more ordered vectors, each associated with its own codebook. This VQ scheme also offers the possibility of being incorporated into the CELP optimization loop by scanning the LSF codebooks for the vectors yielding the lowest overall error between the original and coded speech. This amounts to formant parameter optimization which has so far always been left out of the CELP parameter optimization procedures. This suggestion is applicable to both narrowband and wideband systems. A further enhancement would involve the use of intra-frame interpolation.

The second recommendation deals with reducing the computational load induced by the low–band exhaustive codebook search found in the proposed split–band CELP method. Since the results of Chapter 4 showed that the high–band excitation signal could just be a folded version of the baseband, it could then be possible to carry out the exhaustive low–band codeword search in a narrowband mode (i.e. using a decimated codebook). This would directly reduce the number of computations by half. Once the optimal codeword is found, it would be interpolated and high–band

optimization procedures would be applied to its folded high–band. Another way of reducing the number of computations is to use structured codebooks such as VSELP (Vector Sum Excited Linear Prediction) [20].

The final recommendation proposes means to reduce the number of bits required for some parameters. These can be applied to reduce the operating rate or to increase the parameter update rate to yield higher quality coded speech. First, VQ techniques could be used to code the pitch coefficients when more than one pitch tap is used. Second, some pitch parameters could potentially be updated every second sub-frame, thus reducing their respective transmission rate by half.

Appendix A. Wideband Audio Database

The wideband audio database (Tables A.1 and A.2) contains 48 sentences taken from the Harvard list of phonetically balanced sentences [21]. The sentences have been recorded at a sampling frequency of 16 kHz, and the spectrum information is preserved up to the Nyquist rate (8 kHz). There are four different speakers (2 males, 2 females), and each file identifies the speaker (e.g. F1-09 : Sentence #9 in list of Female #1).

File	Sentence
F1-01	The dark pot hung in the front closet.
F1-02	Carry the pail to the wall and spill it there.
F1-03	The train brought our hero to the big town.
F1-04	Tin cans are absent from store shelves.
F1-05	Slide the box into that empty space.
F1-06	The rude laugh filled the empty room.
F1-07	The plant grew large and green in the window.
F1-08	Tea served from the brown jug is tasty.
F1-09	A dash of pepper spoils beef stew.
F1-10	A zestful food is the hot-cross bun.
F1-11	The cold drizzle will halt the bond drive.
F1-12	The mute muffled the high tones of the horn.
F2-01	He wrote down a long list of items.
F2-02	A siege will crack a strong defense.
F2-03	Grape juice and water mix well.
F2-04	There is a lag between thought and act.
F2-05	Seed is needed to plant the spring corn.
F2-06	The drip of the rain makes a pleasant sound.
F2-07	Draw the chart with heavy black lines.
F2-08	Serve the hot rum to the tired heroes.
F2-09	Much of the story makes good sense.
F2-10	The sun came up to light the eastern sky.
F2-11	The desk was firm on the shaky floor.
F2-12	Nudge gently, but wake her now.

Table A.1 Sentences spoken by females.

File	Sentence
M1-01	The small pup gnawed a hole in the sock.
M1-02	The fish twisted and turned on the bent hook.
M1-03	Press the pants and sew a button on the vest.
M1-04	The swan dive was far short of perfect.
M1-05	The beauty of the view stunned the young boy.
M1-06	Two blue fish swam in the tank.
M1-07	Both lost their lives in the raging storm.
M1-08	The colt reared and threw the tall rider.
M1-09	It snowed, rained and hailed the same morning.
M1-10	Use a pencil to write the first draft.
M1-11	The wrist was badly sprained and hung limp.
M1-12	The frosty air passed thru the coat.
M2-01	The young kid jumped the rusty gate.
M2-02	Guess the results from the first scores.
M2-03	A salt pickle tastes fine with ham.
M2-04	The just claim got the right verdict.
M2-05	These thistles bend in a high wind.
M2-06	Pure bred poodles have curls.
M2-07	Add the store's account to the last cent.
M2-08	The spot on the blotter was made by green ink.
M2-09	Mud was spattered on the front of his white shirt.
M2-10	Fairy tales should be fun to write.
M2-11	The pencils have all been used.
M2-12	Steam hissed from the broken valve.

Table A.2 Sentences spoken by males.

References

1. D. O'Shaughnessy, *Speech Communication, Human and Machine*, Addison-Wesley, 1987.
2. B. Atal and M. R. Schroeder, "Code-excited linear prediction (CELP): high quality speech at very low bit rates," *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing*, pp. 937–940, San Diego, March 1984.
3. N. S. Jayant, "High quality coding of telephone speech and wideband audio," *IEEE Communications Magazine*, pp. 10–20, January 1990.
4. L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, 1978.
5. R. P. Ramachandran and P. Kabal, "Stability and performance analysis of pitch filters in speech coders," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-35, pp. 1419–1426, July 1987.
6. C. K. Un and D. T. Magill, "The residual excited linear predictive vocoder with transmission rate below 9.6 kbits/s," *IEEE Trans. Communications*, vol. COM-23, pp. 1466–1474, December 1975.
7. C. K. Un and J. R. Lee, "On spectral flattening techniques in residual excited linear prediction vocoding," *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing*, pp. 216–219, Paris, May 1982.
8. P. Kabal, "Code excited linear prediction coding of speech at 4.8 Kbits/s," *Rapport technique de l'INRS-Télécommunications*, No. 87-36, July 1987.
9. B. S. Atal and M. R. Shroeder, "Predictive coding of speech signals and subjective error criteria," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-27, No. 3, June 1979, pp. 247–254.
10. Y. Shoham, "Vector predictive quantization of the spectral parameters for low bit rate speech coding," *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing*, pp. 2181–2184, Dallas, April 1987.
11. F. K. Soong and B. H. Juang, "Line spectrum pair (LSP) and speech data compression," *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing*, pp. 1.10.1–1.10.4, San Diego, March 1984.
12. G. S. Kang and L. J. Fransen, "Application of line spectrum pairs to low bit rate speech encoders," *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing*, pp. 7.3.1–7.3.4, Tampa, March 1985.
13. P. Kabal and R. P. Ramachandran, "The computation of line spectral frequencies using Chebyshev polynomials," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-34, pp. 1419–1426, December 1986.
14. N. Sugamura and N. Favardin, "Quantizer design in LSP speech analysis–synthesis," *IEEE Journal on Selected Areas in Communication*, Vol. 6, No. 2, pp.432–440, February 1988.

15. F. K. Soong and B. H. Juang, "Optimal quantization of LSP parameters," *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing*, pp. 394–397, New York, April 1988.

16. J. R. Crosmer and T. P. Barnwell III, "A low bit rate segment vocoder based on line spectrum pairs," *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing*, pp. 7.2.1–7.2.4, Tampa, March 1985.

17. J. L. Moncet and P. Kabal, "Codeword selection for CELP coders," *Rapport technique de l'INRS-Télécommunications*, No. 87-35, July 1987.

18. P. Kabal, J. L. Moncet and C. C. Chu, "Synthesis filter optimization and coding: applications to CELP," *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing*, Section S4.2, pp. 147–150, New York, April 1988.

19. V. Iyengar and P. Kabal, "A low–delay 16 kbits/sec speech coder," *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing*, pp. 243–246, New York, April 1988.

20. Ira A. Gerson and Mark A. Jasiuk, "Vector sum excited linear prediction (VSELP) speech coding at 8 kbps," *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing*, pp. 461–464, Albuquerque, April 1990.

21. "IEEE recommended practice for speech quality measurments," *IEEE Transactions on Audio and Electroacoustices*, Vol. AU-17, No. 3, pp. 225–246, September 1969.